



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, C12N 15/00, C07H 21/04		A1	(11) International Publication Number: WO 97/18333 (43) International Publication Date: 22 May 1997 (22.05.97)									
<p>(21) International Application Number: PCT/US96/18828</p> <p>(22) International Filing Date: 15 November 1996 (15.11.96)</p> <p>(30) Priority Data:</p> <table> <tr> <td>60/006,856</td> <td>16 November 1995 (16.11.95)</td> <td>US</td> </tr> <tr> <td>08/585,758</td> <td>16 January 1996 (16.01.96)</td> <td>US</td> </tr> <tr> <td>08/670,274</td> <td>13 June 1996 (13.06.96)</td> <td>US</td> </tr> </table> <p>(71) Applicant: THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY [US/US]; Suite 350, 900 Welch Road, Palo Alto, CA 94304 (US).</p> <p>(72) Inventors: LI, Limin; Dept. of Genetics/Medicine, Stanford University School of Medicine, Stanford, CA 94305 (US). COHEN, Stanley, N.; Dept. of Genetics/Medicine, Stanford University School of Medicine, Stanford, CA 94305 (US).</p> <p>(74) Agents: BOZICEVIC, Karl et al.; Fish & Richardson P.C., 2200 Sand Hill Road, Menlo Park, CA 94025 (US).</p>		60/006,856	16 November 1995 (16.11.95)	US	08/585,758	16 January 1996 (16.01.96)	US	08/670,274	13 June 1996 (13.06.96)	US	<p>(81) Designated States: AU, CA, JP, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).</p> <p>Published <i>With international search report. Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.</i></p>	
60/006,856	16 November 1995 (16.11.95)	US										
08/585,758	16 January 1996 (16.01.96)	US										
08/670,274	13 June 1996 (13.06.96)	US										
<p>(54) Title: DISRUPTION OF EXPRESSION OF MULTIPLE ALLELES OF MAMMALIAN GENES</p> <p>(57) Abstract</p> <p>Methods are provided for identifying a gene at a random chromosomal locus in the genome of a mammalian cell. The method involves inactivating one copy of the gene by integrating one DNA construct (knockout construct) in that gene copy. The knockout construct includes a positive selection marker region sequence and, in a 5' direction from the selection marker region sequence, a transcription initiation region sequence responsive to a transactivation factor, said transcription initiation region oriented for antisense RNA transcription in the direction away from the selection marker region sequence. The second copy of the gene is inactivated by transforming the cells with a second DNA construct (transactivation construct) containing a gene sequence for the transactivation factor which initiates antisense RNA transcription extending from the knockout construct into the chromosomal locus flanking the knockout construct at its 5' end. Inactivation of both gene copies may result in a change in cell phenotype distinguishable from the wild-type phenotype. Optionally, the wild-type phenotype can be regained by introducing a third construct that can inhibit antisense RNA transcription. A gene is provided associated with tumor susceptibility of mammalian cells, tsg 101.</p>												

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	GB	United Kingdom	MW	Malawi
AT	Austria	GE	Georgia	MX	Mexico
AU	Australia	GN	Guinea	NE	Niger
BB	Barbados	GR	Greece	NL	Netherlands
BE	Belgium	HU	Hungary	NO	Norway
BF	Burkina Faso	IE	Ireland	NZ	New Zealand
BG	Bulgaria	IT	Italy	PL	Poland
BJ	Benin	JP	Japan	PT	Portugal
BR	Brazil	KE	Kenya	RO	Romania
BY	Belarus	KG	Kyrgyzstan	RU	Russian Federation
CA	Canada	KP	Democratic People's Republic of Korea	SD	Sudan
CF	Central African Republic	KR	Republic of Korea	SE	Sweden
CG	Congo	KZ	Kazakhstan	SG	Singapore
CH	Switzerland	LI	Liechtenstein	SI	Slovenia
CI	Côte d'Ivoire	LK	Sri Lanka	SK	Slovakia
CM	Cameroon	LR	Liberia	SN	Senegal
CN	China	LT	Lithuania	SZ	Swaziland
CS	Czechoslovakia	LU	Luxembourg	TD	Chad
CZ	Czech Republic	LV	Latvia	TG	Togo
DE	Germany	MC	Monaco	TJ	Tajikistan
DK	Denmark	MD	Republic of Moldova	TT	Trinidad and Tobago
EE	Estonia	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	UG	Uganda
FI	Finland	MN	Mongolia	US	United States of America
FR	France	MR	Mauritania	UZ	Uzbekistan
GA	Gabon			VN	Viet Nam

Disruption of expression of multiple alleles of mammalian genes**INTRODUCTION****5 Background**

There has been considerable interest in the development of a method for identifying mammalian cell genes whose concurrent homozygous inactivation *de novo* leads to a defined phenotype, where multiple alleles of a gene have been inactivated and where it is easy to confirm that the inactivation results in a phenotype distinguishable from the wild-type. One use of this method is the identification of genes involved in tumor susceptibility.

15 Tumor susceptibility genes may be oncogenes, which are typically upregulated in tumor cells, or tumor suppressor genes, which are down-regulated or absent in tumor cells. Malignancies may arise when a tumor suppressor is lost and/or an oncogene is inappropriately activated. When such mutations occur in somatic cells, they result in the growth of sporadic tumors. Familial predisposition to cancer may occur when there is a mutation, such as loss of an allele encoding a tumor suppressor gene, present in the germline DNA of an individual. In the best characterized familial cancer syndromes, the primary mutation is a loss of function consistent with viability, but resulting in neoplastic change consequent to the acquisition of a second somatic mutation at the same locus.

30 Extensive studies of the early-onset breast cancer families have led to the recent identification of two candidate breast cancer suppressor genes, BRCA1 and BRCA2.

- 2 -

Although frequent mutations of BRCA1 or BRCA2 have been demonstrated in familial early onset breast cancer, this type of cancer represents only about 5-10% of all breast malignancies, and the possible role(s) of BRCA1 and BRCA2 5 in the remaining 90-95% of sporadic breast cancers has not been determined.

Deletion and loss of heterozygosity (LOH) of markers in human chromosome band 11p15 have been shown in a variety of human cancers, including lung cancer, testicular cancer 10 and male germ cell tumor, stomach cancer, Wilms' tumor, ovarian cancer, bladder cancer, myeloid leukemia, malignant astrocytomas and other primitive neuroectodermal tumors, and infantile tumors of adrenal and liver. About 30% of sporadic breast carcinomas show a LOH in this region. 15 Since LOH is believed to indicate inactivation of a tumor suppressor gene at the location where LOH occurs, the frequent LOH found at 11p15 in a variety of human cancers suggests the presence of either a cluster of tumor suppressor genes or a single pleiotropic gene in this 20 region.

The clinical importance of these cancers makes the identification of this putative tumor suppressor gene of great interest for diagnosis, therapy, and drug screening.

Relevant Literature

25 Lemke et al. (1993) Glia 7:263-271 describes loss of function mutations engineered through the expression of antisense RNA from previously cloned genes and through the insertional inactivation of the P_o gene, by homologous recombination in embryonic stem cells, and the generation 30 of P_o -deficient mice. Kamano et al. (1990) Leukemia Res. 10:831-839; van der Krol et al. (1988) Biotechniques 6:958; Katsuki et al. (1988) Science 241:593-595; Owens et al. (1991) Development 112:639-649; and Owens et al. (1991) Neuron 7:565-575 describe changes in cell phenotype

- 3 -

associated with the expression of antisense RNAs in different cell types. Giese et al. (1992) Cell 71:565-576 describes the inactivation of both copies of a gene in a transgenic mouse.

5 Studies of LOH in Wilms' tumors identified a tumor suppressor locus at 11p15, for example see Dowdy et al. (1991) Science 254:293-295. Two familial breast cancer genes have been previously described, BRCA1 in Miki et al. (1994) Science 266:66-71, and BRCA2 in Wooster et al. 10 (1995) Nature 378:789-792.

The interaction of stathmin with a coiled coil domain is described in Sobel (1991) Trends Biochem. Sci. 16:301-305.

SUMMARY OF THE INVENTION

15 Mammalian tumor susceptibility genes and methods for their identification are provided, including the complete nucleotide sequences of human *TSG101* and mouse *tsg101* cDNA. Deletions in *TSG101* are associated with the occurrence of human cancers, for example breast carcinomas. The cancers 20 may be familial, having as a component of risk an inherited genetic predisposition, or may be sporadic. The *TSG101* nucleic acid compositions find use in identifying homologous or related proteins and the DNA sequences encoding such proteins; in producing compositions that 25 modulate the expression or function of the protein; and in studying associated physiological pathways. In addition, modulation of the gene activity *in vivo* is used for prophylactic and therapeutic purposes, such as treatment of cancer, identification of cell type based on expression, 30 and the like. The DNA is further used as a diagnostic for a genetic predisposition to cancer, and to identify specific cancers having mutations in this gene.

The subject invention, in another aspect, provides a method for identifying *de novo* a gene at a random

- 4 -

chromosomal locus of a mammalian cell based on the phenotype produced by interfering with expression of multiple alleles of the gene corresponding to this locus. The method involves inactivating all copies of the gene and 5 any of its alleles which have substantial sequence similarity.

In another aspect, the subject invention provides a rapid method for establishing the function of a gene in a mammalian cell of which at least a portion of the sequence 10 has been previously isolated. In this aspect, the construct integrated in the genome includes two homologous recombination sites which allow for the integration of the construct at the target site. Additionally, the subject invention provides for DNA constructs, vectors, and 15 mammalian cells containing the DNA constructs in their genome.

DESCRIPTION OF THE SPECIFIC EMBODIMENTS

Mammalian *tsg101* gene compositions and methods for their isolation are provided. Of particular interest are 20 the human and mouse homologs. Certain human cancers show deletions at the *TSG101* locus. Many such cancers are sporadic, where the tumor cells have a somatic mutation in *TSG101*. The *TSG101* genes and fragments thereof, encoded protein, and anti-*TSG101* antibodies are useful in the 25 identification of individuals predisposed to development of such cancers, and in characterizing the phenotype of sporadic tumors that are associated with this gene. Tumors may be typed or staged as to the *TSG101* status, e.g. by detection of mutated sequences, antibody detection of 30 abnormal protein products, and functional assays for altered *TSG101* activity. The encoded *TSG101* protein is useful in drug screening for compositions that mimic *TSG101* activity or expression, particularly with respect to *TSG101* function as a tumor suppressor in oncogenesis. *TSG101* can

- 5 -

be used to investigate the interactions with stathmin and the role the complex plays in the regulation of the cell.

The human *TSG101* and mouse *tsg101* gene sequences and isolated nucleic acid compositions are provided. In 5 identifying the human and mouse *TSG101/tsg101* genes, the novel gene discovery approach "random homozygous knock out" was utilized. A retroviral gene search vector carrying a reporter gene was used to select and identify cells containing the vector integrated into target 10 transcriptionally active chromosomal DNA regions, behind chromosomal promoters. 5' to and in reverse orientation to the reporter gene was a regulated promoter with no transcription activity, but which could be highly activated by a transactivator. The system generates large amounts of 15 antisense RNA, which interacts with both alleles of the target gene. Cells transfected with the search vector were further transfected with a plasmid encoding a transactivator. The cells were plated to select for genes whose inactivation led to cellular transformation. While 20 control cell populations formed no colonies in soft agar, the transactivated cells produced 20 colonies. One of these clones was shown to be highly tumorigenic in nude mice. mRNA selection, using a primer specific for the reporter gene, was used to isolate mRNA from the target 25 gene. The mRNA was then used to generate a cDNA clone, which was further used in hybridization screening to isolate the full-length mouse *tsg101* cDNA.

To obtain the human homolog of mouse *tsg101*, the mouse cDNA sequence was used to query dbEST. Ten human 30 partial cDNA sequences included in the database showed 85% to 95% identity to mouse *tsg101*. A conserved sequence was used to design primers that amplify segments of human *TSG101* cDNA, employing total DNA isolated from a human cDNA library as template. The *TSG101* gene has been mapped to 35 human chromosome sub-bands 11p15.1-15.2, and is closely

- 6 -

linked to the Sequence Tagged Site (STS) markers D11S921 through D11S1308 (a detailed map of human genome markers may be found in Dib et al. (1996) Nature 280:152; <http://www.genethon.fr>).

5 The full length human cDNA contains an 1140 bp open reading frame, encoding a 380 amino acid protein. The human and mouse cDNAs are 86% identical at the nucleotide level. The predicted proteins are 94% identical and are distinguished by 20 amino acid mismatches and one gap. A
10 coiled-coil domain (human TSG101 aa 231-302) and a proline-rich domain (human TSG101 aa 130-205, 32% proline) typical of the activation domains of transcription factors are highly conserved between the human and mouse proteins, with only one amino acid mismatch in each of the two
15 domains. The leucine zipper motif in the coiled-coil domain of the human TSG101 protein is identical to the one in the mouse protein.

DNA from a tumor that is suspected of being associated with *TSG101* is analyzed for the presence of an
20 oncogenic mutation in the *TSG101* gene. Sporadic tumors associated with loss of *TSG101* function include a number of carcinomas known to have deletions in the region of human chromosome 11p15, e.g. carcinomas of the breast, lung cancer, testicular cancer and male germ cell tumor, stomach
25 cancer, Wilms' tumor, ovarian cancer, bladder cancer, myeloid leukemia, malignant astrocytomas and other primitive neuroectodermal tumors, and infantile tumors of adrenal and liver.

Characterization of sporadic tumors will generally
30 require analysis of tumor cell DNA, conveniently with a biopsy sample. Where metastasis has occurred, tumor cells may be detected in the blood. Of particular interest is the detection of deletions in the *TSG101* gene, e.g. by amplification of the region and size fractionation of the
35 amplification product; restriction mapping, etc. Screening

- 7 -

of tumors may also be based on the functional or antigenic characteristics of the protein. Immunoassays designed to detect the normal or abnormal TSG101 protein may be used in screening. Alternatively, functional assays, e.g. assays 5 based on detecting changes in the stathmin pathway mediated by TSG101, may be performed.

A wide range of mutations are found, up to and including deletion of the entire short arm of chromosome 11. Specific mutations of interest include any mutation 10 that leads to oncogenesis, including insertions, substitutions and deletions in the coding region sequence, introns that affect splicing, promoter or enhancer that affect the activity and expression of the protein. A "normal" sequence of TSG101 is provided in SEQ ID NO:3 15 (human). In many cases, mutations disrupt the coiled coil domain, resulting in a protein that is truncated or has a deletion in this region. Other mutations of interest may affect the proline rich domain, or other conserved regions of the protein. The leucine zipper within the coiled coil 20 domain is also of particular interest. Biochemical studies may be performed to confirm whether a candidate sequence variation in the TSG101 coding region or control regions is oncogenic. For example, oncogenicity activity of the mutated TSG101 protein may be determined by its ability to 25 complement a loss of TSG101 activity in 3T3 cells, by binding studies with stathmin, etc.

The TSG101 gene may also be used for screening of patients suspected of having a genetic predisposition to TSG101-associated tumors, where the presence of a mutated 30 TSG101 sequence confers an increased susceptibility to cancer. Diagnosis is performed by protein, DNA sequence, PCR screening, or hybridization analysis of any convenient sample from a patient, e.g. biopsy material, blood sample, scrapings from cheek, etc. A typical patient genotype will 35 have an oncogenic mutation on one chromosome. When the

- 8 -

normal copy of *TSG101* is lost, leaving only the reduced function mutant copy, abnormal cell growth is the result.

Prenatal diagnosis may be performed, particularly where there is a family history of the disease, e.g. an 5 affected parent or sibling. A sample of fetal DNA, such as an amniocentesis sample, fetal nucleated or white blood cells isolated from maternal blood, chorionic villus sample, etc. is analyzed for the presence of the predisposing mutation. Alternatively, a protein based 10 assay, e.g. functional assay or immunoassay, is performed on fetal cells known to express *TSG101*.

The DNA sequence encoding *TSG101* may be cDNA or genomic DNA or a fragment thereof. The term "*TSG101 gene*" shall be intended to mean the open reading frame encoding 15 specific *TSG101* polypeptides, as well as adjacent 5' and 3' non-coding nucleotide sequences involved in the regulation of expression, up to about 1 kb beyond the coding region, in either direction. The gene may be introduced into an appropriate vector for extrachromosomal maintenance or for 20 integration into the host.

The term "cDNA" as used herein is intended to include all nucleic acids that share the arrangement of sequence elements found in native mature mRNA species, where sequence elements are exons, 3' and 5' non-coding 25 regions. Normally mRNA species have contiguous exons, with the intervening introns deleted, to create a continuous open reading frame encoding *TSG101*.

The genomic *TSG101* sequence has non-contiguous open reading frames, where introns interrupt the coding regions. 30 A genomic sequence of interest comprises the nucleic acid present between the initiation codon and the stop codon, as defined in the listed sequences, including all of the introns that are normally present in a native chromosome. It may further include the 3' and 5' untranslated regions 35 found in the mature mRNA. It may further include specific

- 9 -

transcriptional and translational regulatory sequences, such as promoters, enhancers, etc., including about 1 kb of flanking genomic DNA at either the 5' or 3' end of the coding region. The genomic DNA may be isolated as a 5 fragment of 50 kbp or smaller; and substantially free of flanking chromosomal sequence.

The nucleic acid compositions of the subject invention encode all or a part of the subject polypeptides. Fragments may be obtained of the DNA sequence by chemically 10 synthesizing oligonucleotides in accordance with conventional methods, by restriction enzyme digestion, by PCR amplification, etc. For the most part, DNA fragments will be of at least 15 nt, usually at least 18 nt, more usually at least about 50 nt. Such small DNA fragments are 15 useful for hybridization screening, etc. Larger DNA fragments, i.e. greater than 100 bp, usually greater than 500 bp, are useful for production of the encoded polypeptide. Single stranded oligonucleotides of from about 18 to 35 nt in length are useful for PCR 20 amplifications. For use in amplification reactions, such as PCR, a pair of primers will be used. The exact composition of the primer sequences is not critical to the invention, but for most applications the primers will hybridize to the subject sequence under stringent 25 conditions, as known in the art. It is preferable to chose a pair of primers that will generate an amplification product of at least about 50 nt, preferably at least about 100 nt. Algorithms for the selection of primer sequences are generally known, and are available in commercial 30 software packages. Amplification primers hybridize to complementary strands of DNA, and will prime towards each other.

The *TSG101* genes are isolated and obtained in substantial purity, generally as other than an intact 35 mammalian chromosome. Usually, the DNA will be obtained

- 10 -

substantially free of other nucleic acid sequences that do not include a *TSG101* sequence or fragment thereof, generally being at least about 50%, usually at least about 90% pure and are typically "recombinant", i.e. flanked by 5 one or more nucleotides with which it is not normally associated on a naturally occurring chromosome.

The DNA sequences are used in a variety of ways. They may be used as probes for identifying other *tsg101* genes. Mammalian homologs have substantial sequence 10 similarity to the subject sequences, i.e. at least 75%, usually at least 90%, more usually at least 95% sequence identity with the nucleotide sequence of the subject DNA sequence. Sequence similarity is calculated based on a reference sequence, which may be a subset of a larger 15 sequence, such as a conserved motif, coding region, flanking region, etc. A reference sequence will usually be at least about 18 nt long, more usually at least about 30 nt long, and may extend to the complete sequence that is being compared. Algorithms for sequence analysis are 20 known in the art, such as BLAST, described in Altschul et al. (1990) J Mol Biol 215:403-10.

Nucleic acids having sequence similarity are detected by hybridization under low stringency conditions, for example, at 50°C and 10XSSC (0.9 M saline/0.09 M sodium 25 citrate) and remain bound when subjected to washing at 55°C in 1XSSC. By using probes, particularly labeled probes of DNA sequences, one can isolate homologous or related genes. The source of homologous genes may be any mammalian species, e.g. primate species; murines, such as rats and 30 mice; canines; felines; bovines; ovines; equines; etc.

The DNA may also be used to identify expression of the gene in a biological specimen. The manner in which one probes cells for the presence of particular nucleotide sequences, as genomic DNA or RNA, is well-established in 35 the literature and does not require elaboration here.

- 11 -

Conveniently, a biological specimen is used as a source of mRNA. The mRNA may be amplified by RT-PCR, using reverse transcriptase to form a complementary DNA strand, followed by polymerase chain reaction amplification using primers 5 specific for the subject DNA sequences. Alternatively, the mRNA sample is separated by gel electrophoresis, transferred to a suitable support, e.g. nitrocellulose, and then probed with a fragment of the subject DNA as a probe. Other techniques may also find use. Detection of mRNA 10 having the subject sequence is indicative of *TSG101* gene expression in the sample.

The subject nucleic acid sequences may be modified for a number of purposes, particularly where they will be used intracellularly, for example, by being joined to a 15 nucleic acid cleaving agent, e.g. a chelated metal ion, such as iron or chromium for cleavage of the gene; as an antisense sequence; or the like. Modifications may include replacing oxygen of the phosphate esters with sulfur or nitrogen, replacing the phosphate with phosphoramido, etc.

20 A number of methods are available for analyzing genomic DNA sequences for the presence of mutations. Where large amounts of DNA are available, the genomic DNA is used directly. Alternatively, the region of interest is cloned into a suitable vector and grown in sufficient quantity for 25 analysis, or amplified by conventional techniques, such as the polymerase chain reaction (PCR). The use of the polymerase chain reaction is described in Saiki et al. (1985) Science 239:487, and a review of current techniques may be found in Sambrook, et al. Molecular Cloning: A 30 Laboratory Manual, CSH Press 1989, pp.14.2-14.33.

PCR is particularly useful for detection of oncogenic mutations. In many cases such mutations involve a deletion at the *TSG101* locus. For example, primers specific for *TSG101* are used to amplify all or part of the 35 gene. The amplification products are then analyzed for

- 12 -

size, where a deletion will result in a smaller than expected product. Where the deletion is very large, there may be a complete absence of the specific amplification product. Alternatively, analysis may be performed on mRNA 5 from a cell sample, where the RNA is converted to cDNA, and then amplified (RT-PCR).

A detectable label may be included in the amplification reaction. Suitable labels include fluorochromes, e.g. fluorescein isothiocyanate (FITC), 10 rhodamine, Texas Red, phycoerythrin, allophycocyanin, 6-carboxyfluorescein (6-FAM), 2',7'-dimethoxy-4',5'-dichloro-6-carboxyfluorescein (JOE), 6-carboxy-X-rhodamine (ROX), 6-carboxy-2',4',7',4,7-hexachlorofluorescein (HEX), 15 5-carboxyfluorescein (5-FAM) or N,N,N',N'-tetramethyl-6-carboxyrhodamine (TAMRA), radioactive labels, e.g. ^{32}P , ^{35}S , ^3H ; etc. The label may be a two stage system, where the amplified DNA is conjugated to biotin, haptens, etc. having a high affinity binding partner, e.g. avidin, specific antibodies, etc., where the binding partner is conjugated 20 to a detectable label. The label may be conjugated to one or both of the primers. Alternatively, the pool of nucleotides used in the amplification is labeled, so as to incorporate the label into the amplification product.

The amplified or cloned fragment may be sequenced by 25 dideoxy or other methods, and the sequence of bases compared to the normal *TSG101* sequence. Hybridization with the variant, oncogenic sequence may also be used to determine its presence, by Southern blots, dot blots, etc. Single strand conformational polymorphism (SSCP) analysis, 30 denaturing gradient gel electrophoresis (DGGE), and heteroduplex analysis in gel matrices are used to detect conformational changes created by DNA sequence variation as alterations in electrophoretic mobility. The hybridization pattern of a control and variant sequence to an array of 35 oligonucleotide probes immobilised on a solid support, as

- 13 -

described in WO 95/11995, may also be used as a means of detecting the presence of variant sequences. Alternatively, where an oncogenic mutation creates or destroys a recognition site for a restriction endonuclease, 5 the fragment is digested with that endonuclease, and the products size fractionated to determine whether the fragment was digested. Fractionation is performed by gel electrophoresis, particularly acrylamide or agarose gels.

The subject nucleic acids can be used to generate 10 transgenic animals or site specific gene modifications in cell lines. The modified cells or animals are useful in the study of *TSG101* function and regulation. For example, a series of small deletions and/or substitutions may be made in the *TSG101* gene to determine the role of different 15 exons in oncogenesis, signal transduction, etc. One may also provide for expression of the *TSG101* gene or variants thereof in cells or tissues where it is not normally expressed or at abnormal times of development. In addition, by providing expression of *TSG101* protein in 20 cells in which it is otherwise not normally produced, one can induce changes in cell behavior.

DNA constructs for homologous recombination will comprise at least a portion of the *TSG101* gene with the desired genetic modification, and will include regions of 25 homology to the target locus. Alternatively, constructs may that do not target to the native locus, but integrate at random sites int he genome. Conveniently, markers for positive and negative selection are included. Methods for generating cells having targeted gene modifications through 30 recombination are known in the art. For various techniques for transfecting mammalian cells, see Keown et al. (1990) Methods in Enzymology 185:527-537.

For embryonic stem (ES) cells, an ES cell line may be employed, or ES cells may be obtained freshly from a 35 host, e.g. mouse, rat, guinea pig, etc. Such cells are

- 14 -

grown on an appropriate fibroblast-feeder layer or grown in the presence of leukemia inhibiting factor (LIF). When ES cells have been transformed, they may be used to produce transgenic animals. After transformation, the cells are 5 plated onto a feeder layer in an appropriate medium. Cells containing the construct may be detected by employing a selective medium. After sufficient time for colonies to grow, they are picked and analyzed for the occurrence of homologous recombination. Those colonies that show 10 homologous recombination may then be used for embryo manipulation and blastocyst injection. Blastocysts are obtained from 4 to 6 week old superovulated females. The ES cells are trypsinized, and the modified cells are injected into the blastocoel of the blastocyst. After 15 injection, the blastocysts are returned to each uterine horn of pseudopregnant females. Females are then allowed to go to term and the resulting litters screened for mutant cells having the construct. By providing for a different phenotype of the blastocyst and the ES cells, chimeric 20 progeny can be readily detected.

The chimeric animals are screened for the presence of the modified gene and males and females having the modification are mated to produce homozygous progeny. If the gene alterations cause lethality at some point in 25 development, tissues or organs can be maintained as allogeneic or congenic grafts or transplants, or in *in vitro* culture. The transgenic animals may be any non-human mammal, such as laboratory animals, domestic animals, etc. The transgenic animals may be used in functional studies, 30 drug screening, etc., e.g. to determine the effect of a candidate drug on tumor cells.

The subject gene may be employed for producing all or portions of the TSG101 protein. Peptides of interest include the coiled-coil domain (aa 231-302) and the 35 proline- rich domain (aa 130-205). For expression, an

- 15 -

expression cassette may be employed, providing for a transcriptional and translational initiation region, which may be inducible or constitutive, the coding region under the transcriptional control of the transcriptional 5 initiation region, and a transcriptional and translational termination region. Various transcriptional initiation regions may be employed which are functional in the expression host.

The peptide may be expressed in prokaryotes or 10 eukaryotes in accordance with conventional ways, depending upon the purpose for expression. For large scale production of the protein, a unicellular organism or cells of a higher organism, e.g. eukaryotes such as vertebrates, particularly mammals, may be used as the expression host, 15 such as *E. coli*, *B. subtilis*, *S. cerevisiae*, and the like. In many situations, it may be desirable to express the TSG101 gene in a mammalian host, whereby the TSG101 protein will be glycosylated.

With the availability of the protein in large 20 amounts by employing an expression host, the protein may be isolated and purified in accordance with conventional ways. A lysate may be prepared of the expression host and the lysate purified using HPLC, exclusion chromatography, gel electrophoresis, affinity chromatography, or other 25 purification technique. The purified protein will generally be at least about 80% pure, preferably at least about 90% pure, and may be up to and including 100% pure. By pure is intended free of other proteins, as well as cellular debris.

30 TSG101 polypeptides are useful in the investigation of the stathmin signaling pathway, which is involved in the regulation and relay of diverse signals associated with cell growth and differentiation. The coiled coil domain of TSG101 interacts with stathmin. The structure of TSG101 35 indicates that it is a transcription factor, which may act

- 16 -

as a downstream effector of stathmin signaling. The normal and mutated forms of TSG101 polypeptides may be used for binding assays with other proteins, to detect changes in phosphorylation, etc. that may affect this pathway. Yeast 5 has been shown to be a powerful tool for studying protein-protein interactions through the two hybrid system described in Chien et al. (1991) P.N.A.S. **88**:9578-9582.

Binding assays of TSG101 to DNA may be performed in accordance with conventional techniques for DNA 10 footprinting, to determine the sequence motifs that are recognized by TSG101. In vitro transcription assays may be used, to determine how complexes comprising polymerase and transcriptional activation factors are affected by the presence of TSG101.

15 The polypeptide is used for the production of antibodies, where short fragments provide for antibodies specific for the particular polypeptide, whereas larger fragments or the entire gene allow for the production of antibodies over the surface of the polypeptide or protein. 20 Antibodies may be raised to the normal or mutated forms of TSG101. The coiled coil, leucine zipper and proline rich domains of the protein are of interest as epitopes, particularly to raise antibodies that recognize common changes found in oncogenic TSG101. Antibodies may be 25 raised to isolated peptides corresponding to these domains, or to the native protein. Antibodies that recognize TSG101 are useful in diagnosis, typing and staging of human tumors, e.g. breast carcinomas.

Antibodies are prepared in accordance with 30 conventional ways, where the expressed polypeptide or protein may be used as an immunogen, by itself or conjugated to known immunogenic carriers, e.g. KLH, pre-S HBsAg, other viral or eukaryotic proteins, or the like. Various adjuvants may be employed, with a series of 35 injections, as appropriate. For monoclonal antibodies,

after one or more booster injections, the spleen may be isolated, the splenocytes immortalized, and then screened for high affinity antibody binding. The immortalized cells, e.g. hybridomas, producing the desired antibodies 5 may then be expanded. For further description, see Monoclonal Antibodies: A Laboratory Manual, Harlow and Lane eds., Cold Spring Harbor Laboratories, Cold Spring Harbor, New York, 1988. If desired, the mRNA encoding the heavy and light chains may be isolated and mutagenized by cloning 10 in *E. coli*, and the heavy and light chains may be mixed to further enhance the affinity of the antibody.

The antibodies find particular use in diagnostic assays for carcinomas and other tumors associated with mutations in TSG101. Staging, detection and typing of 15 tumors may utilize a quantitative immunoassay for the presence or absence of normal TSG101. Alternatively, the presence of mutated forms of TSG101 may be determined. A reduction in normal TSG101 and/or presence of abnormal TSG101 is indicative that the tumor is TSG101-associated.

20 A sample is taken from a patient suspected of having a TSG101-associated tumor. Samples, as used herein, include biological fluids such as blood, cerebrospinal fluid, tears, saliva, lymph, dialysis fluid and the like; organ or tissue culture derived fluids; and fluids 25 extracted from physiological tissues. Also included in the term are derivatives and fractions of such fluids. Biopsy samples are of particular interest, e.g. carcinoma samples, organ tissue fragments, etc. Where metastasis is suspected, blood samples may be preferred. The number of 30 cells in a sample will generally be at least about 10^3 , usually at least 10^4 more usually at least about 10^5 . Usually a lysate of the cells is prepared.

35 Diagnosis may be performed by a number of methods. The different methods all determine the absence or presence of normal or abnormal TSG101 in patient cells suspected of

- 18 -

having a mutation in TSG101. For example, detection may utilize staining of histological sections, performed in accordance with conventional methods. The antibodies of interest are added to the cell sample, and incubated for a 5 period of time sufficient to allow binding to the epitope, usually at least about 10 minutes. The antibody may be labeled with radioisotopes, enzymes, fluorescers, chemiluminescers, or other labels for direct detection. Alternatively, a second stage antibody or reagent is used 10 to amplify the signal. Such reagents are well-known in the art. For example, the primary antibody may be conjugated to biotin, with horseradish peroxidase-conjugated avidin added as a second stage reagent. Final detection uses a substrate that undergoes a color change in the presence of 15 the peroxidase. The absence or presence of antibody binding may be determined by various methods, including microscopy, spectrophometry, scintillation counting, etc.

An alternative method for diagnosis depends on the in vitro detection of binding between antibodies and TSG101 20 in a lysate. Measuring the concentration of TSG101 binding in a sample or fraction thereof may be accomplished by a variety of specific assays. A conventional sandwich type assay may be used. For example, a sandwich assay may first attach TSG101-specific antibodies to an insoluble surface 25 or support. The particular manner of binding is not crucial so long as it is compatible with the reagents and overall methods of the invention. They may be bound to the plates covalently or non-covalently, preferably non-covalently.

30 The insoluble supports may be any compositions to which polypeptides can be bound, which is readily separated from soluble material, and which is otherwise compatible with the overall method. The surface of such supports may be solid or porous and of any convenient shape. Examples 35 of suitable insoluble supports to which the receptor is

- 19 -

bound include beads, e.g. magnetic beads, membranes and microtiter plates. These are typically made of glass, plastic (e.g. polystyrene), polysaccharides, nylon or nitrocellulose. Microtiter plates are especially 5 convenient because a large number of assays can be carried out simultaneously, using small amounts of reagents and samples.

Patient sample lysates are then added to separately assayable supports (for example, separate wells of a 10 microtiter plate) containing antibodies. Preferably, a series of standards, containing known concentrations of normal and/or abnormal TSG101 is assayed in parallel with the samples or aliquots thereof to serve as controls. Preferably, each sample and standard will be added to 15 multiple wells so that mean values can be obtained for each. The incubation time should be sufficient for binding, generally, from about 0.1 to 3 hr is sufficient. After incubation, the insoluble support is generally washed of non-bound components. Generally, a dilute non-ionic 20 detergent medium at an appropriate pH, generally 7-8, is used as a wash medium. From one to six washes may be employed, with sufficient volume to thoroughly wash non-specifically bound proteins present in the sample.

After washing, a solution containing a second 25 antibody is applied. The antibody will bind TSG101 with sufficient specificity such that it can be distinguished from other components present. The second antibodies may be labeled to facilitate direct, or indirect quantification of binding. Examples of labels that permit direct 30 measurement of second receptor binding include radiolabels, such as ^3H or ^{125}I , fluorescers, dyes, beads, chemiluminescers, colloidal particles, and the like. Examples of labels which permit indirect measurement of binding include enzymes where the substrate may provide for 35 a colored or fluorescent product. In a preferred

- 20 -

embodiment, the antibodies are labeled with a covalently bound enzyme capable of providing a detectable product signal after addition of suitable substrate. Examples of suitable enzymes for use in conjugates include horseradish 5 peroxidase, alkaline phosphatase, malate dehydrogenase and the like. Where not commercially available, such antibody-enzyme conjugates are readily produced by techniques known to those skilled in the art. The incubation time should be sufficient for the labeled ligand to bind available 10 molecules. Generally, from about 0.1 to 3 hr is sufficient, usually 1 hr sufficing.

After the second binding step, the insoluble support is again washed free of non-specifically bound material. The signal produced by the bound conjugate is detected by 15 conventional means. Where an enzyme conjugate is used, an appropriate enzyme substrate is provided so a detectable product is formed.

Other immunoassays are known in the art and may find use as diagnostics. Ouchterlony plates provide a simple 20 determination of antibody binding. Western blots may be performed on protein gels or protein spots on filters, using a detection system specific for TSG101 as desired, conveniently using a labeling method as described for the sandwich assay.

25 By providing for the production of large amounts of TSG101 protein, one can identify ligands or substrates that bind to, modulate or mimic the action of TSG101. Areas of investigation include the development of cancer treatments. Drug screening identifies agents that provide a replacement 30 for TSG101 function in abnormal cells. The role of TSG101 as a tumor suppressor indicates that agents which mimic its function will inhibit the process of oncogenesis. Of particular interest are screening assays for agents that have a low toxicity for human cells. A wide variety of 35 assays may be used for this purpose, including labeled in

- 21 -

vitro protein-protein binding assays, electrophoretic mobility shift assays, immunoassays for protein binding, and the like. The purified protein may also be used for determination of three-dimensional crystal structure, which can be used for modeling intermolecular interactions, transcriptional regulation function, etc.

The term "agent" as used herein describes any molecule, protein, or pharmaceutical with the capability of altering or mimicking the physiological function of TSG101.

10 Generally a plurality of assay mixtures are run in parallel with different agent concentrations to obtain a differential response to the various concentrations. Typically, one of these concentrations serves as a negative control, i.e. at zero concentration or below the level of

15 detection.

Candidate agents encompass numerous chemical classes, though typically they are organic molecules, preferably small organic compounds having a molecular weight of more than 50 and less than about 2,500 daltons.

20 Candidate agents comprise functional groups necessary for structural interaction with proteins, particularly hydrogen bonding, and typically include at least an amine, carbonyl, hydroxyl or carboxyl group, preferably at least two of the functional chemical groups. The candidate agents often

25 comprise cyclical carbon or heterocyclic structures and/or aromatic or polyaromatic structures substituted with one or more of the above functional groups. Candidate agents are also found among biomolecules including peptides, saccharides, fatty acids, steroids, purines, pyrimidines,

30 derivatives, structural analogs or combinations thereof.

Candidate agents are obtained from a wide variety of sources including libraries of synthetic or natural compounds. For example, numerous means are available for random and directed synthesis of a wide variety of organic

35 compounds and biomolecules, including expression of

- 22 -

randomized oligonucleotides and oligopeptides. Alternatively, libraries of natural compounds in the form of bacterial, fungal, plant and animal extracts are available or readily produced. Additionally, natural or 5 synthetically produced libraries and compounds are readily modified through conventional chemical, physical and biochemical means, and may be used to produce combinatorial libraries. Known pharmacological agents may be subjected to directed or random chemical modifications, such as 10 acylation, alkylation, esterification, amidification, etc. to produce structural analogs.

Where the screening assay is a binding assay, one or more of the molecules may be joined to a label, where the label can directly or indirectly provide a detectable 15 signal. Various labels include radioisotopes, fluorescers, chemiluminescers, enzymes, specific binding molecules, particles, e.g. magnetic particles, and the like. Specific binding molecules include pairs, such as biotin and streptavidin, digoxin and antidigoxin etc. For the 20 specific binding members, the complementary member would normally be labeled with a molecule that provides for detection, in accordance with known procedures.

A variety of other reagents may be included in the screening assay. These include reagents like salts, 25 neutral proteins, e.g. albumin, detergents, etc that are used to facilitate optimal protein-protein binding and/or reduce non-specific or background interactions. Reagents that improve the efficiency of the assay, such as protease inhibitors, nuclease inhibitors, anti-microbial agents, 30 etc. may be used. The mixture of components are added in any order that provides for the requisite binding. Incubations are performed at any suitable temperature, typically between 4 and 40°C. Incubation periods are selected for optimum activity, but may also be optimized to

- 23 -

facilitate rapid high-throughput screening. Typically between 0.1 and 1 hours will be sufficient.

Other assays of interest detect agents that mimic TSG101 function. For example, candidate agents are added 5 to a cell that lacks functional TSG101, and screened for the ability to reproduce TSG101 function, e.g. prevent growth of 3T3 cells in soft agar.

The compounds having the desired pharmacological activity may be administered in a physiologically 10 acceptable carrier to a host for treatment of cancer attributable to a defect in *tsg101* function. The inhibitory agents may be administered in a variety of ways, orally, topically, parenterally e.g. subcutaneously, intraperitoneally, intravascularly, etc. Topical 15 treatments are of particular interest. Depending upon the manner of introduction, the compounds may be formulated in a variety of ways. The concentration of therapeutically active compound in the formulation may vary from about 0.1-100 wt.%.

20 The pharmaceutical compositions can be prepared in various forms, such as granules, tablets, pills, suppositories, capsules, suspensions, salves, lotions and the like. Pharmaceutical grade organic or inorganic carriers and/or diluents suitable for oral and topical use 25 can be used to make up compositions containing the therapeutically-active compounds. Diluents known to the art include aqueous media, vegetable and animal oils and fats. Stabilizing agents, wetting and emulsifying agents, salts for varying the osmotic pressure or buffers for 30 securing an adequate pH value, and skin penetration enhancers can be used as auxiliary agents.

The gene may also be used for gene therapy. Vectors useful for introduction of the gene include plasmids and viral vectors. Of particular interest are retroviral-based 35 vectors, e.g. moloney murine leukemia virus and modified

- 24 -

human immunodeficiency virus; adenovirus vectors, etc. Gene therapy may be used to treat cancerous lesions, an affected fetus, etc., by transfection of the normal gene into suitable cells. A wide variety of viral vectors can 5 be employed for transfection and stable integration of the gene into the genome of the cells. Alternatively, micro-injection may be employed, fusion, or the like for introduction of genes into a host cell. See, for example, Dhawan et al. (1991) Science 254:1509-1512 and Smith et al. 10 (1990) Molecular and Cellular Biology 3268-3271.

Concurrent Disruption of Genes

The subject invention also provides a method for identifying a gene at a random chromosomal locus of a mammalian cell on the basis of a phenotype produced by 15 homozygous inactivation of gene function. The method includes the concurrent inactivation of other copies of the genes or alleles of the gene that are substantially similar in their DNA sequence, followed by a determination of whether inactivation of the gene and its alleles has 20 resulted in a cell phenotype distinguishable from the wild-type phenotype. Additionally, the chromosomal locus containing the inactivated gene may be identified.

One copy of the gene is inactivated by the integration of a DNA construct at a random or unselected 25 chromosomal locus, or one that is selected for its proximity to an expressed gene. The construct integrated at the random chromosomal locus contains a transcription initiation region sequence responsive to a transactivation factor. Transcription occurs in the opposite direction to 30 a coding region for a promoterless positive selection marker gene. Hereinafter, the transactivation factor responsive transcriptional initiation region will be referred to as the "TF promoter." The positive selection marker gene carried by the construct is expressed only when

- 25 -

the construct is integrated 3' in relation to an endogenous gene promoter, with the endogenous promoter directing transcription toward the positive selection marker gene. Expression of the positive selection marker gene may also 5 require that it be in the correct reading frame to express an active positive selection marker, either fused or non-fused to the expression product of a portion of the endogenous gene at the locus of integration.

Additionally, either the TF promoter is activated by 10 a factor which can be added to the medium or by a transactivation factor encoded and expressed by a second construct, which second construct is introduced into the host. This factor serves to activate antisense RNA transcription from the TF promoter. Antisense RNA 15 transcription from the TF promoter extends in the 3' direction relative to the orientation of the TF promoter, and into the flanking chromosomal locus 5' to the insertion of the marker gene. Binding or hybridization of the antisense RNA transcription product initiated from the TF 20 promoter inhibits expression of the other alleles which have substantially similar sequence to the chromosomal DNA sequence transcribed from the TF promoter, and sequences complementary to the antisense RNA. (By "similar sequence" is intended sufficient similarity to bind to a messenger 25 RNA sequence to provide an observable change in the function of the messenger RNA sequence). In this way, the production of all of the same or allelic proteins, including enzyme isoforms, may be prevented.

By first introducing the construct with the 30 promoterless marker gene, one can select for cells in which the marker gene is being expressed, so that there is a high likelihood that the construct is positioned in an actively transcribed gene. By expanding these cells, one may then use these cells for introduction of the expression 35 construct containing the transactivation factor gene. This

- 26 -

expression construct will have its own marker gene, so that one can select for those cells which have the transactivation factor expression construct. The resulting cells should, for the most part, be cells which transcribe 5 a transcription product antisense to the construct encoded by the gene containing the inserted construct, which results in inhibition of expression of proteins by all alleles of the locus at which the promoterless marker gene integrated.

10 The cells which have been selected for the markers associated with the two constructs may now be screened to determine whether gene inactivation has resulted in a specifically desired cell phenotype distinguishable from the wild-type phenotype, or alternatively selection of 15 cells expressing the desired phenotype can be carried out. Additionally, the chromosomal locus flanking the 5' end of the integrated construct may be identified.

As a further check on whether the observed change in phenotype is as a result of the transcription of an 20 antisense strand, the transcription of the antisense strand can be reversed by introducing a third construct into cells with a modified phenotype. This construct comprises an expression construct expressing a recombinase gene sequence that would serve to excise the transactivation gene or the 25 TF promoter, where flanking consensus sequences recognized by the recombinase gene are present in the transactivation and/or TF promoter constructs, resulting in the excision of these elements, and such other portions of the construct as are deemed appropriate. Alternatively, turn off of the 30 antisense promoter can be accomplished by using a promoter that is regulated by hormones, chemical agents, temperature, or other agents.

The method includes the preparation and introduction of a gene search construct that includes a TF promoter and 35 a construct that expresses a transactivator protein,

- 27 -

normally the introduction being sequential, into a mammalian cell. (By "gene search construct" is intended a promoterless reporter gene and 5' sequences that may generate fusion transcripts originating in DNA 5' to the 5 reporter gene, so that the fusion construct will include any portion of the coding region of the endogenous gene between the endogenous promoter and the 5' sequence of the construct, the 5' sequence of the construct and also all or functional part of the reporter gene. By "transactivator 10 construct" is intended a DNA molecule which comprises a transcriptional initiation region, a translational initiation sequence, a coding sequence encoding a protein that activates the TF promoter, and a translational and transcriptional termination region. All of the regions and 15 sequences will be functional in the host of interest). Preferably, a first construct ("gene search construct") comprising a promoterless marker gene and a transactivation factor responsive promoter ("TF promoter"), directing transcription in the opposite direction of the coding 20 region of the promoterless marker gene, is introduced first into the cells and the cells expanded and selected for cells which express the promoterless marker gene. At this time, the TF promoter is inactive and will not interfere with the transcription of the promoterless marker gene,

25 A second construct may also be introduced which expresses a marker gene and a transactivation factor which acts on the TF promoter to activate the promoter, resulting in transcription of an antisense strand. As transcription of the antisense strand extends into the chromosomal DNA 30 flanking the gene search construct, it serves to inactivate other genes having substantially similar sequences to the antisense strand of flanking DNA. Those cells having the gene search construct, with the first construct at a locus which provides for expression of the promoterless marker

- 28 -

gene, will under conditions of transactivation be screened or selected for change in phenotype.

Rather than activate the TF promoter with a transactivation factor from expression of the expression 5 construct, one may employ a TF promoter which is responsive to agents, e.g., compounds or other stimuli, which may be added to the medium or provided as a change in environment, e.g. heat. There are many promoters which have responsive elements, e.g., tetracycline or hormonal, such as steroid, 10 responsive elements, where compounds can be added to the medium which will turn on the TF promoter, e.g., tetracycline derivatives or hormones, such as glucocorticoids.

To further establish whether the change in phenotype 15 is as a result of the production of an antisense sequence, one may provide for reversal of the transcription of the antisense sequence, by providing at least one of the constructs with sequences that result in excision of the DNA region between the excision sequences. One introduces 20 a third construct, a recombinase expression construct, into the host comprising a marker gene and a gene encoding a recombinase which acts on the excision sequences for excision. The cells may then be screened for reversal of the phenotype, indicating that the phenotype was the result 25 of the production of the antisense sequence. Alternatively, expression of the antisense promoter or of a transactivator protein that turns on this promoter can be regulated using one or more of the agents indicated above and below.

30 The transcriptional initiation region of the TF promoter or of the transactivator gene employed in this invention may be varied widely, as required by the particular application. The transcriptional initiation region of the TF promoter or transactivator gene may be 35 constitutive or inducible, as appropriate, may include

enhancers, repressors, or other regulatory sequences, which may be regulated in cis or trans. Regulation may also be as a result of additives to the media, e.g., tetracycline or hormones, e.g., glucocorticoids. The initiation region may 5 be from any source, where the initiation region is functional in the host. Thus, the initiation region may be from structural genes of the host, from viral genes functional in the host, or combinations of such promoter regions, or synthetic promoter regions, as appropriate.

10 The promoter region may be a single promoter region (associated with a single gene) or a combination of 5' regions associated with different genes. The promoter region will usually be chosen to provide the desired level of transcription for the particular coding region or gene.

15 Promoters which may find use in mammalian cells include SV40 promoters, glucocorticoid inducible promoters, CMV promoters, β -actin promoters, etc. The coding region or gene will be under the transcriptional and translational regulation of the initiation region. There will also be a

20 translational and transcriptional termination region downstream in the direction of transcription from the gene. Since, for the most part, the termination region is not important to the functioning of expression, a wide variety of termination regions may be used from a wide variety of

25 host genes, viral genes functional in the host, or the like.

For producing the antisense RNA, a promoter is employed which will not be active in the cell, except in conjunction with a transactivation factor or other inducing 30 agent or condition necessary to activate the promoter. (See the above discussion concerning inducible promoters). This factor will be necessary for transcription initiation and can be supplied by a second construct introduced into the host cell or by adding the appropriate agent or providing 35 the appropriate condition(s) for activation of the TF

- 30 -

promoter. With the expression construct which will not be required if induction of transcription does not require a transactivator protein, one can ensure strong binding of the transactivation factor to the TF promoter, desirably 5 using a chimeric protein which combines a DNA-binding domain and a transcription factor, which directly or indirectly binds to RNA Polymerase II. The DNA-binding domain binds to a DNA sequence which desirably is not found in mammalian cells, and therefore would not be expected to 10 bind at endogenous mammalian promoters. The DNA-binding domain allows for potent binding of the transactivation factor to a unique DNA sequence while orienting the transcription factor close to an RNA polymerase II binding site for interaction of the transcription factor with the 15 transcriptional machinery of the host mammalian cell.

The marker gene of the expression construct expressing the transactivator may be a single gene or a fused gene comprising two different markers which may be selected differentially. Single gene markers include *neo*, 20 which can provide for G418 resistance. Combination genes may include *lacZ* and aminoglycoside phosphotransferase (*aph*) which provides G418 resistance, hygromycin resistance (*hyg*) and the herpes simplex thymidine kinase (TK) gene, which provides resistance to hygromycin and sensitivity to 25 ganciclovir.

The DNA-binding domain of the transactivator protein will usually be from a host foreign to the target host, usually a unicellular microorganism, insect, plant or the like, so as to be unlikely to be recognized by DNA binding 30 proteins in the host, while the transcriptional activation domains will be from the host or other source which provides a factor which binds to the target host RNA polymerase II. Conveniently, the DNA-binding domain is derived from a DNA-binding protein isolated from bacterial 35 cells and the transcription activation domain is derived

- 31 -

from a transcription activation domain that binds to RNA polymerase II from a common genus, e.g., mammalian cells for a mammalian host.

In a preferred embodiment, the transactivation factor contains the DNA-binding domain of the lac repressor at its amino terminus and the transcription activation domain from the herpes simplex virus virion protein 16 (VP16) at its carboxyl terminus, or the like.

The first DNA construct comprising the promoterless marker gene may also be referred to as the "knockout" construct. This knockout construct includes the TF promoter which comprises the DNA sequence bound by the transactivation factor. The expression construct comprising the transactivation factor gene or second construct could be part of the first construct or be introduced first, but this would not allow cells to be selected for appropriate integration at a gene locus and then subsequent expansion, without the complication of the antisense RNA also being produced, which might interfere with expression of the marker. It is therefore desirable that the second construct be introduced after cells having integration of the first construct downstream from a promoter have been selected, unless the transactivation factor gene is inducible, so that transcription of the transactivation gene may be initiated after selection for integration of the knockout construct.

The TF promoter typically includes a region consisting of sequence repeats, two or more, usually at least about 5 and not more than 20, and in a preferred embodiment, 14, which are tightly bound by the DNA-binding domain of the transactivation factor. Additionally, the TF promoter includes a promoter sequence for binding the RNA polymerase II of the host cell to place RNA polymerase II in close relationship with the transcription initiation domain. In a preferred embodiment the promoter sequence

responsive to the transactivation factor consists of a minimal SV-40 promoter, which lacks the enhancer sequences and GC-rich sequences typically found in the SV-40 early transcription promoter, but which can still bind the RNA 5 polymerase, and spaced sets of *lac* operators located upstream from the promoter.

The *lac* operator sequences cause strong binding of the *lac* repressor DNA-binding domain at the transcription initiation region. The minimal SV-40 promoter binds the 10 RNA polymerase II, which is also bound by the transcription factor of the chimeric protein to enhance transcription from the TF promoter.

The transcription initiation region is oriented so that RNA transcription initiated from the transcription 15 initiation region extends into the chromosomal locus flanking the knockout construct at its 5' end to provide for the antisense transcript on the non-coding strand, so as to be complementary to the sense strand from which the mRNA is transcribed.

20 The knockout construct also includes a coding region sequence for a positive selection marker. The term "coding region sequence" ordinarily refers to the coding region for a polypeptide without a promoter. The coding region sequence is located so as to allow for fusion of the coding 25 region sequence with an exon of the gene into which the gene search construct is integrated, usually upstream (5' direction) in relationship to the direction of transcription of the TF promoter. The coding region sequence is conveniently downstream of the TF promoter, 30 where a splice site may be employed which is positioned to remove the TF promoter when the coding region sequence is spliced to a chromosomal exon. Since the positive selection marker coding region sequence lacks a promoter and the transactivatable antisense promoter region sequence is 35 oriented in the opposite direction, the selection marker

- 33 -

coding region sequence is only expressed if the knockout construct is integrated downstream from an endogenous gene promoter. Additionally, when integrated within a translated portion of the gene, the coding region needs to 5 be in the correct reading frame to form a fused protein consisting of an active positive selection marker and the truncated polypeptide of the gene. This can be done by including 5' to the positive selection marker a splice acceptor sequence in one or more translational reading 10 frames.

Optionally, the knockout construct contains a splice acceptor sequence which is located usually about 20 or fewer base pairs upstream of the positive selection marker region, although it may be within or downstream from the TF 15 promoter. The splice acceptor sequence is useful in case the knockout construct has integrated at an intron or 3' UTR of a chromosomal gene and is employed for splicing the precursor RNA to incorporate the positive selection marker gene sequence in the mRNA. If the coding sequence is 20 incorporated into the 5' UTR, the coding sequence will include an initiation codon.

The knockout construct with the 3'-splice site may be organized with the TF promoter upstream or downstream from the 3'-splice site (if downstream, desirably the TF 25 promoter sequence will lack a stop codon in phase with the coding region sequence); the 3'-splice site; and the coding region sequence.

To obtain integration one may introduce the bare DNA into the host cell. Preferably a construct will be used 30 which enhances integration, such as an integrating virus, e.g., self-inactivating Moloney Murine Leukemia Virus, adenovirus, transposons and a transposase, etc. Alternatively integrating of DNA can be accomplished following introduction of naked DNA by electroporation. 35 Depending upon the particular vector employed for

- 34 -

integration, the integration may be more or less random, depending on whether the vector has sequence preferences. It should be noted that retroviral insertions have some preference for actively transcribed regions, so that there 5 will be some enrichment for integration into genes with retroviral integration sequences.

Stable maintenance of the first and second constructs in the mammalian cell and integration of the knockout construct at a random chromosomal locus as 10 described above, where the promoterless marker gene is placed under the transcriptional construct of an endogenous promoter, results in (i) expression of the positive selection marker coding region sequence, and (ii) in conjunction with transactivation factor expression or 15 availability, binding of the factor to the transcription initiation region sequence activating antisense RNA transcription extending into the 5' region of the gene locus in relation to the TF promoter.

One copy of the gene is inactivated by the insertion 20 of the knockout construct in its sequence. Expression of other copies of genes similar, particularly complementary, to the antisense transcript is inhibited. The antisense RNA forms duplexes with cellular RNAs that are similar to the antisense RNA transcript. Duplex formation inhibits the 25 function of such cellular RNAs. Such inactivation of the related genes may cause a change in the phenotype of the cells.

In order to investigate whether a change in phenotype is associated with the antisense RNA production, 30 one provides for reversal of the production of the RNA antisense. Optionally, the first and/or second constructs may contain two site-specific recombination sites delimiting the regions associated with the production of the antisense sequence. That is, they may delimit the TF 35 promoter and/or the transactivation gene sequence.

.35.

Preferably, these sites delimit the transactivation gene sequence or functional portion thereof, e.g., promoter or coding region. In order to provide for excision of the sequence(s) associated with the production of the antisense 5 RNA, a third DNA construct (recombinase expression construct) containing a recombinase gene may be prepared and integrated into the host comprising the first and second constructs. A marker gene may also be provided as part of the construct, so as to select for cells into which 10 the recombinase construct has integrated. These selected cells may be expanded and screened for change of phenotype. Alternatively, where the transactivation factor is provided extrinsic to the cells, the medium or environment may be changed to stop transcription of the antisense sequence. 15 Another alternative is to regulate the antisense (TF) promoter by use of hormones, Tc, etc.

If regulation by removal of the transactivator is desired, the sequence between the consensus sequences for deletion comprises a negative selection marker, 20 particularly in conjunction with the gene expressing the transactivation factor. Expression of the recombinase gene typically causes the excision of a fragment of the transactivation factor construct's DNA which contains the gene for the transactivation factor and the gene for the 25 negative selection marker. The cells may then be selected for lack of expression of the negative selection marker.

The recombinase is an enzyme which causes the excision of any DNA sequence that is delimited by site-specific recombination sites. These site-specific 30 recombination sites are sequences typically between 20 to 100 base pairs, usually about 30 to 50 base pairs and are exogenous to the host cell. Site-specific recombination sites contain two recombinase recognition sequences in inverted orientation at an overlap region. The 35 recombination sites are oriented as repeats to cause

- 36 -

segment excision. The recombinase may be a member of the integrase protein family which includes cre protein, int protein and FLP protein.

Expression of the recombinase in a mammalian cell 5 causes excision of the DNA fragment which is delimited by the two site-specific recombination sites and which will desirably include the transactivation factor coding region sequence. The transactivation factor is therefore no longer expressed, and antisense RNA transcription is no 10 longer initiated from the knockout construct. Typically, the cell phenotype, where there are two or more copies of the gene, should revert to the wild-type phenotype, where the previous change in phenotype was as a result of the production of the antisense sequence.

15 All the DNA constructs contain selection marker gene sequences for monitoring insertion of the constructs in mammalian cells comprising the different construct genes and for allowing for selection of such cells substantially free of other cells not comprising the construct sequences.

20 The positive selection marker gene is a gene sequence that allows for selection of target cells in which the subject constructs have been introduced. Positive selection marker genes include the neo gene for resistance to G418, the hygromycin resistance gene, and the like. A negative 25 selection marker gene is typically the herpes simplex virus thymidine kinase (tk) gene, whose expression can be detected by the use of nucleoside analogs, such as acyclovir or gancyclovir, for their cytotoxic effects on cells that contain a functional tk gene.

30 As discussed above, the knockout construct contains a coding region sequence for a positive selection marker to select for cells expressing the positive selection marker. The positive selection marker is only expressed if it is downstream in relationship to an endogenous promoter and,

- 37 -

when fused to a coding region, is fused in frame with the native protein.

The transactivation factor construct contains a positive selection marker gene to select for cells 5 expressing the positive selection marker. The transactivation construct may also include a negative selection marker gene in the DNA fragment that is excised from the construct in the presence of a recombinase. In this manner the excision of the DNA fragment containing the 10 transactivation factor gene sequence is monitored.

Additionally, the constructs may contain other sequences required for manipulation of the constructs. For example, restriction sites are necessary for manipulating the sequences in the constructs. Other sequences which may 15 be present include primer initiation sequences for amplifying DNA, origins for cloning, markers for cloning hosts, sequences aiding in integration into the host chromosome, and the like.

The constructs may be included in vectors for 20 introducing the constructs into mammalian cells. When the vectors are introduced into a cell by retroviral infection, these sequences include long terminal repeats and packaging signals. When the introduced vectors are to replicate episomally in a mammalian cell, the vectors include a viral 25 origin of replication.

According to the subject invention, both knockout and transactivator constructs ordinarily are introduced into the target mammalian cells. Particularly the mammalian cells are mouse cells, rat cells, primate cells, 30 e.g., sequentially human cells, rabbit cells or the like. Other eukaryotic hosts may also be used, such as plant cells, insect cells, fish cells, fungal cells, and the like. The mammalian cells may be normal cells, in a differentiated or undifferentiated state, e.g., stem cells. 35 Alternatively, the cells may be transfected with naked DNA.

- 38 -

Desirably, the cells are maintainable in culture and allow for the introduction of new genetic material.

The constructs may be introduced into the target cell in accordance with known ways. For example, the 5 constructs may be introduced by retroviral infection, electroporation, fusion, polybrene, lipofection, calcium phosphate precipitated DNA, or other conventional techniques. Particularly, the knockout construct is introduced by viral infection for largely random 10 integration of the construct in the genome. The transactivation construct is introduced into cells by any of the methods described above. After introduction of each construct into target mammalian cells, the cells are grown in a selective medium to select for cells that express the 15 appropriate selection markers, substantially free of cells that do not express the selection markers. For example, cells receiving a knockout construct containing the neomycin coding region sequence are grown in a medium containing G418, and cells receiving a transactivation 20 construct containing the hygromycin resistance gene sequence are grown in a medium containing hygromycin.

Stable expression of the first positive selection marker coding sequence indicates that the knockout construct has been integrated into a chromosomal locus, 25 downstream of an endogenous promoter. Stable expression of the second positive selection marker gene sequence indicates that the transactivation construct has been stably introduced in the cells.

The cells that have received the knockout construct 30 and stably express both positive selection markers are assayed for a cell phenotype distinguishable from the wild-type phenotype. Different types of phenotypes may include changes in growth pattern and requirements, sensitivity or resistance to infectious agents or chemical substances, 35 changes in the ability to differentiate or nature of the

- 34 -

differentiation, changes in morphology, changes in response to changes in the environment, e.g., physical changes or chemical changes, changes in response to genetic modifications, and the like.

5 For example, the change in cell phenotype may be the change from normal cell growth to uncontrolled cell growth. The cells may be screened by any convenient assay which provides for detection of uncontrolled cell growth. One assay which may be used is a methylcellulose assay with
10 bromodeoxyuridine (BrdU). Another assay which is effective is the use of growth in agar (0.3 to 0.5% thickening agent). A test for tumorigenicity may also be used, where the cells may be introduced into a susceptible host, e.g. immunosuppressed, and the formation of tumors determined.

15 Alternatively, the change in cell phenotype may be the change from a normal metabolic state to an abnormal metabolic state. In this case, cells are assayed for their metabolite requirement, such as amino acids, sugars, cofactors, or the like, for growth. Initially, about 10
20 different metabolites may be screened at a time to assay for utilization of the different metabolites. Once a group of metabolites has been identified that allows for cell growth, where in the absence of such metabolites the cells do not grow, the metabolites are screened individually to
25 identify which metabolite is assimilable or essential.

Alternatively, the change in cell phenotype may be a change in the structure of the cell. In such a case, cells might be visually inspected under a light or electron microscope.

30 The change in cell phenotype may be a change in the differentiation program of a cell. For example, the differentiation of myoblasts to adult muscle fibers can be investigated. The differentiation of myoblasts can be induced by an appropriate change in the growth medium and
35 can be monitored by determining the expression of specific

- 40 -

polypeptides, such as myosin and troponin, which are expressed at high levels in adult muscle fibers.

The change in cell phenotype may be a change in the commitment of a cell to a specific differentiation program. 5 For example, cells derived from the neural crest, if exposed to glucocorticoids, commit to becoming adrenal chromaffin cells. However, if the cells are exposed instead to fibroblast growth factor or nerve growth factor, the cells eventually become sympathetic adrenergic neuronal 10 cells. If the adrenergic neuronal cells are further exposed to ciliary neurotrophic factor or to leukemia inhibitory factor, the cells become cholinergic neuronal cells. Cells transfected by the method of the subject invention can therefore be exposed to either 15 glucocorticoids or any of the factors, and changes in the commitment of the cells to the different differentiation pathways can be monitored by assaying for the expression of polypeptides associated with the various cell types.

After establishing a change in phenotype, the 20 chromosomal region flanking the knockout construct DNA may be identified using PCR with the construct sequence as a primer for unidirectional PCR, or in conjunction with a degenerate primer, for bidirectional PCR. The sequence may then be used to probe a cDNA or chromosomal library for the 25 locus, so that the region may be isolated and sequenced. Alternatively, the region knocked out by antisense RNA may be sequenced and, if a large enough portion is identified, the coding region may be used in the sense direction and a polypeptide sequence obtained. The resulting peptide may 30 then be used for the production of antibodies to isolate the particular protein. Also, the peptide may be sequenced and the peptide sequence compared with known peptide sequences to determine any homologies with other known polypeptides. Various techniques may be used for 35 identification of the gene at the locus and the protein

- 41 -

expressed by the gene, since the subject methodology provides for a marker at the locus, obtaining a sequence which can be used as a probe and, in some instances, for expression of a protein fragment for production of 5 antibodies. If desired the protein may be prepared and purified for further characterization.

The subject method, in another aspect, is employed to identify the function of a gene when at least part of the sequence of the gene is known. The method includes the 10 inactivation of both gene copies to determine a change in cell phenotype, or a loss of function, associated with the inactivation of specific alleles of the gene.

The method includes the preparation of a knockout construct including both the promoter region sequence and 15 the positive selection marker coding region sequence. Additionally, the construct contains two homologous recombination sites delimiting the promoter region sequence and the positive selection marker sequence. These homologous recombination sites are homologous to sequences 20 of the known gene and allow for insertion of the knockout construct sequence flanked by the two recombination sites into the known gene.

These homologous recombination sites will typically be not more than about 2 kbp, usually not more than 1 kbp. 25 The sites will typically be not less than 0.05 kbp, usually not less than 0.1 kbp. The regions of homology between these recombination site sequences and the target sequences will typically be at least about 90%, usually greater than 95%. The regions of homology are preferably within coding 30 regions, such as exons, of the gene.

Additionally, the knockout construct may contain the transactivation factor gene sequence, so that no other construct is required for performing the subject invention.

Using this approach, one may inhibit expression of 35 the alleles of a gene, where only a partial sequence is

- 42 -

known and determine whether the expression product has an effect on phenotype, since all of the copies of the gene and related alleles may be inhibited from expression. In this manner, without knowing what the gene is, one may 5 conveniently determine whether the function of the gene is of interest.

The following examples are offered by way of illustration and not by way of limitation.

EXPERIMENTAL

10

Example 1

The method described below allows for the identification and isolation of new genes involved in the regulation of cell growth and differentiation. Preparation of constructs, methods for mammalian cell transformation, 15 assays for uncontrolled cell growth, and methods for identifying the new gene are provided.

Results

Experimental Approach and Construction of Gene Search Vectors. pLLGSV, a retroviral gene search vector 20 derived from self-inactivating Moloney murine leukemia virus (MLV) (Hawley et al., *PNAS USA* (1987) 84:2406-2410; Brenner et al., *PNAS USA* (1989) 86:5517-5521) carries the β -geo (Friedrich and Soriano, *Genes & Develop.* (1991) 5:1513-1523) reporter gene. This reporter, a fusion of the 25 *E. coli lacZ* and aminoglycoside phosphotransferase (*aph* or "neo") genes, encodes resistance to the antibiotic G418, which was used to select and identify cells containing virus integrated into transcriptionally active chromosomal DNA regions behind chromosomal promoters. An adenovirus- 30 derived splice acceptor (Friedrich and Soriano, 1991 *supra*) was inserted at the 5' end of β -geo to enhance the fusion of β -geo mRNA to upstream transcripts encoded by chromosomally-encoded exons. 5' to, and in reverse

- 43 -

orientation to β -geo, is a regulated promoter formed by fusion of the SV40 early T antigen minimal promoter sequence to 14 *E.coli lacZ* operators (Labow et al., *Mol. Cell. Biol.* (1990) 10:3343-3356); this promoter has no transcription activity, but can be highly activated in trans by a transactivator, Lap348 (Labow et al., 1990, *supra*), containing the operator-binding domain of the *E. coli lacI* repressor and the herpes simplex virus transactivation domain VP16. The system was designed to generate large amounts of antisense RNA, which interact not only with the sense RNA encoded by the allele with the integrated gene search vector, but also with the sense RNA encoded by other allele(s) of the same gene.

pLLGAV was first transfected into helper cells (GP+E-86) to generate infectious viruses to infect NIH3T3 cells. A population of G418 resistant NIH3T3 cells, containing the pLLGSV vector integrated at transcriptionally active sites behind chromosomal promoters throughout the 3T3 cell genome, were transfected with transactivator vector pLLTX. pLLTX encodes both the Lap348 and HyTK, a fusion of a hygromycin resistance (*hyg*) gene and the herpes simplex virus thymidine kinase (TK) gene (Lupton et al., *Mol. Cell. Biol.* (1991) 11:3374-3378). Transfectants expressing HyTK are resistant to *hyg* but sensitive to gancyclovir (*gcv*), which specifically kills cells expressing herpes TK. In contrast, in the absence of HyTK expression, cells are *hyg*-sensitive and *gcv*-resistant. Two *lox* sites from bacteriophage P1 flanking the transactivator and HyTK genes allow excision of the Lap348/HyTK segment from chromosomes of cells by Cre, a *lox*-specific recombinase (Sauer and Henderson, *Nature* (1989) 298:447-451) expressed from pRSV-cre introduced into *hyg* resistant cells by electroporation. Cells in which the Lap348/HyTK segment has been excised, and in which the

- 44 -

regulated promoter consequently has been turned off, are detected by their resistance to gcv.

hyg resistant NIH3T3 cells were plated in 0.5% agarose to select for transformation phenotype, i.e., to 5 select genes whose inactivation may contribute to cellular transformation. Excision of LAP348 from transformed cells by Cre generated transactivator deleted clones. Comparing the phenotypes of the cells with transactivator present and cells with transactivator deleted, further confirms that 10 cellular transformation results from transactivator generated antisense RNA. Cells with transactivator deleted can be used for cloning of the gene containing the gene search vector.

Isolation of Clones Showing Transformed Phenotype.

15 2.5 x 10⁸ NIH 3T3 cells were infected with viral supernatant from a culture of a PLLGSV-transfected helper cell clone selected for its ability to produce a high titer of infectious virus. Infected cells containing chromosomally integrated PLLGSV were either selected on plates for G418 20 resistance or collected by fluorescence-activated cell sorting (Brenner et al., 1989, *supra*) for β -galactosidase activity; the cell population obtained by either method showed variable degrees of deep blue staining by X-gal. A pool of more than 5 x 10⁶ clones containing retroviral 25 integrations selected for G418 resistance was transfected with the transactivator vector PLLTX by electroporation; colonies selected for hyg resistance were pooled and plated in 0.5% agarose. Whereas no cells in a similarly-sized uninfected NIH 3T3 population formed colonies on this 30 concentration of agarose, the PLLGSV infected population produced 20 colonies. One of these clones, SL6 was expanded into cell line, which was transfected with pRSV-cre to generate cells with deleted transactivator (SL6 Δ T cells. Both SL6 and SL6 Δ T cells were injected into nude mice 35 subcutaneously, where only SL6 cells were highly

- 45 -

tumorigenic. Although SL6ΔT cells produced a small tumor in one mouse, neither control NIH3T3 cells nor NIH3T3 cells transfected with pLLTX alone produced any tumor. Only SL6 cells produced spontaneous metastases to the lung.

5 Replating of SL6, SL6ΔT and control cells into 0.5% agarose showed that only SL6 cells formed large colonies. To examine the regulation of reporter gene expression by transactivator, SL6 and SL6ΔT cells were assayed for β -galactosidase activity (Table 1). When transactivator was

10 present in SL6 cells, expression of reporter gene was almost complete by shut off, compared to background control cells; when transactivator was removed by cre-lox recombination in SL6ΔT cells, the reporter gene was highly expressed. These results indicate that transactivator

15 generated antisense RNA can effectively inactivate gene expression.

Table 1. Characterization of SL6

Transactivator	3T3	3T3	SL6	SL6
	-	+	-	+
β -Galactosidase Activity (U/ μ g)	9.26 ^a	10.05	1225.80	19.88
20 Growth in 0.5% Agarose	-	-	$20/10^5$ ^b	$1000/10^5$
Tumorigenicity in Nude Mice	0/10	0/10	1/10	10/10
Spontaneous Lung Metastasis ^c	0/10	0/10	0/10	8/10

^aMeans of triplicates.

^bThe colonies formed by SL6 without transactivator were significantly 25 smaller than those formed by SL6 with transactivator.

^cMice were sacrificed at day 32 with lung metastases were confirmed by histology.

A genomic southern blot of SL6 cells using an 1.3 kb neo fragment probe showed a single chromosomal integration 30 of PLLGSV; both the reporter gene and the regulated promoter were faithfully duplicated in accordance with the retroviral life cycle. Northern blotting of poly(A) RNA

- 46 -

isolated from SL6ΔT using a 550 bp fragment of 5' β -geo as a probe, showed a major transcript of 7 Kb in length, and two transcripts of 7.5 Kb and 6.5 Kb in smaller amount. Hybridization with the cloned gene confirmed that the 7 Kb and 6.5 Kb transcripts were fusion transcripts of the reporter gene and mRNA initiated at a chromosomally-located promoter external to the vector. During cDNA cloning (see below), we also isolated many alternatively spliced cDNA products, in which the splice acceptor site of the second 10 copy of the reporter gene in the provirus had been spliced to several cryptic splice donors of the first reporter gene, and such aberrant splicing may result in multiple transcripts in Northern blots, as has been observed previously (Friedrich and Soriano, 1991, *supra*).

15 *cDNA Cloning and Sequence Analysis.* A biotin labeled oligodeoxyribonucleotide that corresponds to the 5' end of β -geo was used to select β -geo fusion mRNA from SL6ΔT cells by hybridization; the hybridized mRNAs were purified using streptavidin-coated paramagnetic particles, 20 reverse transcribed, converted to double strand cDNA, cloned into the *E. coli* plasmid pAmp1, and sequenced by standard methods. The cloned 120 bp cDNA segment contained 70 bp of a novel sequence fused in frame to the splice acceptor site 5' to β -geo. A data base search using the 25 BLAST program (Altschul et al., *J. Mol. Biol.* (1990) 215:403-410) showed 97% identity to a mouse partial cDNA sequence of unknown function identified by its expression during differentiation of F9 mouse embryonal carcinoma cells (Nishiguchi et al., (1994) *J. Bio. Chem.* 116:128-139.

30 A mouse NIH 3T3 cell cDNA library was screened with the 70 bp cDNA probe to obtain a full length gene. Four positive clones were isolated, and all contained a 1148 bp open translational reading frame (ORF) encoding a predicted 381 amino acid protein of 43,108 kDa. The gene defined by 35 this sequence was designated as tumor susceptibility gene

- 47 -

101 (*tsg101*). A potential consensus sequence for initiation of translation, followed by an adenosine residue three bases upstream of a putative ATG translation start codon, was located near the 5' end of the *tsg101*. A splice 5 donor consensus sequence (AG) was observed 72 nucleotides into the cDNA sequence analyzed and four codons downstream of the ATG.

The sequence of full length *tsg101* cDNA and the predicted amino acid sequence of the Tsg101 protein were 10 used to search the non-redundant DNA and protein sequence databases of the National Center for Biotechnology Information using the BLAST program. This analysis indicated that amino acids 231 to 301 of *tsg101* are identical, except for two mismatches to cc2, an α -helix 15 domain encoded by a partial cDNA clone identified by its ability to express a protein that interacts with stathmin (Maucuer et al., *PNAS USA* (1995) 92:3100-3104); an evolutionarily-conserved phosphoprotein implicated in the integration and relay of diverse signals regulating cell 20 growth (Sobel, *Trends Biochem. Sci.* (1991) 16:301-305). The algorithm of Stock and colleagues (Lupas et al., *Science* (1991) 252:1162-1164) predicts with a probability of ~99.8% that the helical domain of Tsg101 will form a coiled-coil structure. A protein pattern search of full 25 length Tsg101 identified a leucine zipper domain within the coiled-coil domain of Tsg101, consistent with the observed ability of the cc2 domain to interact with stathmin. Additionally, seven potential protein kinase C phosphorylation sites (aa11, 38, 85, 88, 215, 225, 357), 30 five potential Casein kinase II phosphorylation sites (aa38, 210, 249, 265, 290), two potential N-myristylation sites (aa55, 156), and three potential N-glycosylation sites (aa44, 150, 297) were present in Tsg101 (Bairoch and Bucher, *Nucleic Acids Res.* (1994) 22:3583-9). 35 A protein motif search (Prints, Leads University, UK)

- 48 -

showed that aa37-46 of Tsg101 resembles the helix-turn-helix signature domain of the bacteriophage 1 repressor (i.e., HTHLAMBDA) (Brennan and Matthews, *J. Biol. Chem.* (1989) 264:1903-1906), and that aa73-83 resembles a fungal 5 Zn-cys bi-nuclear cluster signature (FUNGALZCYS) (Pan and Coleman, *PNAS USA* (1990) 87:2077-2081).

Expression of tsg101 Sense and Antisense RNA Cause Transformation of Naive NIH3T3 Cells. To confirm the role of *tsg101* in cell growth, we investigated the effects of 10 overexpression of *tsg101* in sense and antisense orientations in naive NIH 3T3 cells. In both instances, the *tsg101* sequence was expressed in stably transfected cells under control of the cytomegalovirus (CMV) promoter. Expression of *tsg101* in either the sense or antisense 15 orientation resulted in transformation of naive NIH3T3 cells, as indicated by the ability to form colonies on 0.5% agarose. Whereas no colonies were observed in cells transfected with the vector lacking the insert or in mock transfected cells.

20 Experimental Procedures

Construction of Vectors. To construct the self-inactivated retroviral gene search vector pLLGSV, a 4.3 kb *Xho*I-*Xho*I fragment from pSA β -geo (Friedrich and Soriano, *Genes & Develop.* (1991) 5:1513-1523), containing β -geo 25 reporter gene and a splice acceptor sequence 5' to the reporter, was ligated into a *Xho*I linker site of pACYC184 plasmid (Chang and Cohen, *J. Bacteriol.* (1978) 134:1141-1156) that had been digested with *Tth*111I and *Xba*I. The *Nhe*I site of pACYC was then deleted and the *Xho*I site 5' to 30 the β -geo reporter gene was converted into a *Nhe*I site by linker insertion; a 1.45 kb *Pvu*II-*Stu*I fragment containing 14 *lac* operator repeats and a SV40 minimal promoter sequence from pL14CAT (Labow et al., 1990, *supra*) was introduced into an *Spe*I 5' to the splice acceptor site and

- 49 -

β -geo in the opposite orientation to β -geo. The polyadenylation signal of β -geo was deleted by *Xba*I digestion and replaced with a *Nhe*I linker. This 5.4 kb *Nhe*I-*Nhe*I fragment was then ligated in the same orientation 5 as retroviral transcription, into a *Nhe*I site at the deleted 3' LTR of pHHAM (Hawley et al., *PNAS USA* (1987) 84:2406-2410) after *Nhe*I partial digestion.

The transactivator vector pLLTX was derived from pHCMVLAP348 (Labow et al., *Mol. Cell. Biol.* (1990) 10:3343-10 3356). The *Hind*III site at the 3' end of the transactivator was first deleted and a 1952 bp *Sfi*I fragment containing a HyTK gene expression cassette (Lupton et al., *Mol. Cell. Biol.* (1991) 11:3374-3378), was ligated 15 into the *Hind*III site upstream of transactivator to yield pLAPHyTK. A 200 bp DNA fragment containing two directly repeated *loxP* sites derived from pBS30 (Sauer and Henderson, *Nucleic Acids Res.* (1989) 17:147-161) was introduced into a *Clal* site of pLAPHyTK to give pLLTX. 20 pBS30 was first digested with *Sal*I and *Bam*HI, and ligated with a *Hind*III linker; then the vector was digested with *Aat*II and *Xho*I to generate this 200 bp fragment with two directly repeated *loxP* sites. This 200 bp fragment was 25 ligated into a *Clal* site of pLAPHyTK to give pLLTX.

To construct the expression vector pLLEXP I, a 1410 bp fragment [containing a human β -actin promoter, the puromycin resistance gene *pac*, and an SV40 poly(A) site] 25 was first cloned into the *Bam*H1 site of pBR332 to generate pBR- β -*pac*. The *Sfi*I fragment containing the HyTK gene expression cassette (Lupton et al., 1991, *supra*) was then 30 inserted into a *Bam*HI site of pBR- β -*pac*, after *Bam*HI partial digestion to give pBR- β -*pac*-HyTK. The expression vector pLLEXP I was generated by *Nhe*I and *Bgl*III digestion of pBR- β -*pac*-HyTK to remove the HyTK gene and replaced by cDNA inserts.

- 50 -

Cell Culture and Transfection. NIH 3T3 cells (ATCC) and GP+E-86 cells (Markowitz et al., *J. Virol.* (1988) 62:1120-1124) were cultured in Dubecco's modified Eagle's medium (DMEM) supplemented with 10% calf serum (3T3) or 10% new born calf serum (GP+E-86), 100U/ml penicillin, and 100 mg/ml streptomycin. DNA transfection was carried out by electroporation (Potter et al., *PNAS USA* (1984) 81:7161-7165) using Cell-Porator Electroporation systems I (Life Technologies, Inc.) and Lipofectamin (Life Technologies, Inc.) according to the protocol of the manufacturer.

Retroviral Infection of Mouse Fibroblast NIH3T3 Cells. To generate infectious retrovirus, pLLGSV was linearized by treatment with *Scal*I and transfected into helper cell line GP+E-86 by electroporation. The transfected GP+E-86 cells were replated on day 3 and selected with 800 μ g/ml G418 for 2-3 weeks. All G418 resistant clones were isolated and expanded in 24-well plates. Culture supernatant from each clone was incubated with NIH 3T3 cells in the presence of polybrene (8 μ g/ml) for 8 hr, and the frequency of integration behind the chromosomal promoter was subsequently determined by X-gal staining of the infected NIH 3T3 cells. The helper cell clones giving the highest frequency of integrations behind chromosomal promoters were expanded and culture supernatant was collected for large scale infection of NIH 3T3 cells.

Isolation of Transformed Clones and Tumorigenicity Assay. Cultures of G418 resistant NIH 3T3 cells were trypsinized and transfected with *Hind*III linearized pLLTX DNA by electroporation. The transfected cells were selected with 500 μ g/ml of hygromycin for 12-18 days. All hygromycin resistant clones were plated into 0.5% agarose (Li et al., *J. Natl. Cancer Inst.* (1989) 81:1406-1412), 4 to 6 weeks later, the colonies formed in 0.5% agarose were isolated and expanded to cell lines. To assay the tumorigenicity of the transfected cells, 10^5 cells were

- 51 -

injected into nude mice (NIH *nu/nu*, female and 6 weeks of age) subcutaneously over the lateral thorax. The animals were examined twice weekly and sacrificed five weeks later. The neoplastic nature of local tumors and lung metastases 5 were confirmed by histologic examination (Fidler, *Cancer Metastasis Rev.* (1986) 5:29-49).

cDNA Cloning and Screening of cDNA Library. A biotin labeled oligodeoxyribonucleotide (27 mer) that corresponds to the 5' end of the β -geo reporter gene was 10 hybridized with polyadenylated mRNA from SL6AT cells, and captured with Streptavidin paramagnetic particles (Promega). The oligo-hybridized mRNA was eluted and reverse transcribed with a gene specific primer corresponding to a sequence located upstream of the biotin 15 labeled oligo into first strands of cDNA. A uracil DNA glycosylase (UDG) cloning site (Booth et al., *Gene* (1994) 146:303-308) was incorporated into the gene specific primer to facilitate cDNA cloning. The first strand cDNA was then 3' tailed with (dG)_n by terminal transferase, and converted 20 into ds cDNA using a UDG-oligo d(c)₂₀ primer and DNA polymerase. The dscDNAs were cloned into the UDG-cloning vector pAMP1 (Life Technologies, Inc.) and screened for fusion to β -geo. A 70 bp cDNA segment of novel sequence fused in frame to the splice acceptor site 5' to β -geo was 25 used as a probe to screen a mouse NIH 3T3 cDNA library (Stratagene). Positive clones were sequenced with Sequenase 2.0 (USB) for both strands.

Southern and Northern Blot Analysis. Genomic DNA was isolated by standard procedure. Total RNA was isolated 30 with RNA STAT-60 (TEL-TEST), and poly(A) mRNA was isolated with PolyATtract (Promega). Both DNA and RNA blots were probed with PCR generated single-stranded DNA probes.

- 52 -

Example 2.

Chromosomal mapping studies assigned *TSG101* to human chromosome 11 band p15, a region showing loss of heterozygosity primarily in breast cancer but also in other 5 human malignancies, and proposed previously to contain tumor suppressor gene(s). Intragenic deletions in *TSG101* were identified in six of fourteen metastatic breast cancer cell lines, and six of fourteen primary human breast carcinomas had mutations in the gene. All of these 10 mutations directly removed all or part of the cc2 domain. These findings support the conclusion that *TSG101* is a suppressor of abnormal cell growth and additionally demonstrate that this gene has an important role in human breast cancer.

15 Results

Cloning and Characterization of Human TSG101 cDNA. *tsg101* was initially identified in mouse cells by a novel gene discovery approach that enables regulated functional inactivation of multiple copies of previously unknown genes 20 and selection for cells that show a phenotype resulting from such inactivation. To obtain *TSG101*, the human homolog of mouse *tsg101*, the 1448 bp mouse cDNA sequence was used to query dbEST of the National Cancer for Biotechnology Information (NCBI) by the BLAST program. Ten 25 human partial cDNA sequences (Expressed Sequences Tags, EST) included in the database showed 85% to 95% identity to mouse *tsg101* cDNA. A 27 bp sequence contained within a region of 100% identity between ESTs H53754 and Z30135 was used to design the UDG primers Pa-UDG and Pd-UDG; these 30 primers plus two other UDG primers (Pb-UDG and Pc-UDG) corresponding to sequences bracketing the vector cloning site of a lgt10-based human cDNA library were used to amplify by PCR the 5' (Pc-UDG and Pd-UDG) and 3' (Pa-UDG and Pb-UDG) segments of human *TSG101* cDNA, employing total DNA

- 53 -

isolated from the human cDNA library as template. The longest 5' and 3' PCR products were then joined in the UDG cloning vector pAMP1.

A 1494 bp cloned human cDNA insert was termed full length *TSG101* cDNA. Sequence analysis of this cDNA identified a 1140 bp open reading frame predicted to encode a 380 amino acid protein with a molecular mass of 42.841 kDa and a pI of 5.87. The human and mouse cDNAs are 86% identical at the nucleotide level. The predicted proteins are 94% identical and are distinguished by 20 amino acid mismatches and one gap. A coiled-coil domain (human *TSG101* aa 231-302) and a proline-rich domain (human *TSG101* aa 130-205, 32% proline) typical of the activation domains of transcription factors are highly conserved between the human and mouse proteins, with only one amino acid mismatch in each of the two domains. The leucine zipper motif in the coiled-coil domain of the human *TSG101* protein is identical to the one in the mouse protein. Other conserved features identified in human *TSG101* include seven putative protein kinase C phosphorylation sites (aa 11, 38, 86, 89, 215, 225, 357), five potential case in kinase II phosphorylation sites (aa 38, 210, 249, 265, 290) and three potential N-glycosylation sites (aa 44, 150, 297). Analysis of the human *TSG101* cDNA and protein sequences by the BLAST program search of NCBI database did not reveal any significant homology with the sequences for any other human genes.

Expression of *TSG101* in human tissues was examined on a multiple-tissue Northern blot probed with full length *tsg101* cDNA. A single 1.5 kb transcript was observed in all eight human tissues tested and was slightly more prominent in RNA isolated from liver and pancreas. The size of this transcript indicates that the 1494 bp cDNA corresponds to full length native *TSG101* mRNA.

Chromosomal localization of human and mouse *TSG101* genes. By using PCR primers that specifically amplify a human *TSG101* sequence from the 3'-untranslated region, genomic DNA from a panel of 18 human x Chinese hamster 5 hybrid cell lines was analyzed. The expected 210 bp PCR product was obtained only from hybrid cell lines that had retained human chromosome 11 and from total human genomic DNA, but not from hamster DNA. The human-specific PCR product was also generated from a cell line (31-2A HAT) 10 that retained only the short arm of chromosome 11 (11p), whereas no PCR amplification was observed using the same primers in a cell line that had only the long arm of chromosome 11 (11q). By concordant segregation and by excluding all other chromosomes, the human *TSG101* gene is 15 assigned to chromosome arm 11p.

To obtain a human *TSG101* genomic DNA probe suitable for mapping by fluorescence *in situ* hybridization (FISH), the same set of PCR primers employed for the analysis of hybrid cell lines was used to screen a PAC library 20 containing human genomic DNA inserts. Two overlapping clones, PAC1 and PAC2, each containing ~150 kb inserts, were isolated and confirmed to contain *TSG101* human genomic DNA by Southern blotting using a 5' human *TSG101* cDNA fragment as probe. Fluorescence *in situ* hybridization of 25 the two PAC clones to human chromosome spreads gave identical results, which confirmed the localization of *TSG101* on chromosome arm 11p by our somatic cell hybrid analysis. A fluorescence signal on both chromatids of both copies of chromosome 11 was seen in 20 metaphase cells 30 analyzed. Based on the chromosomal R-banding pattern, *TSG101* is assigned to chromosome 11 bands p15.1-p15.2.

Radiation hybrid (RH) mapping provides another independent approach to map human genes and to position them relative to polymorphic markers on the linkage map. 35 PCR typing for human *TSG101* of the Stanford G3 human RH

- 55 -

mapping panel revealed a positive result in 11 of the 83 RH cell lines (retention frequency 13.25%). By two point linkage analysis *TSG101* was found to be closely linked to Sequence Tagged Site (STS) markers D11S921, D11S899, and 5 D11S1308. Both D11S921 and D11S1308 are on the Whitehead Institute integrated map and radiation hybrid map and their physical positions approximately correspond to 11p15.

To map *tsg101* in the mouse, a mapping panel of 22 mouse x rodent hybrid cell lines was analyzed by PCR using 10 mouse gene-specific primers. The presence or absence of mouse chromosome 7 in hybrid cell lines was in complete concordance with the 202 bp mouse *tsg101* PCR product. All other mouse chromosomes were excluded by at least 3 discordant hybrids. An attempt to place the gene on the 15 mouse linkage map by typing an interspecies backcross panel was not successful, as no difference between C57BL/6 and *M. spretus* patterns were detectable by single strand conformational analysis (SSCA) of PCR products. Given the known conserved syntenic regions on human chromosome 11p 20 and mouse chromosome 7, the mapping of the mouse gene provides further evidence that the human and mouse sequences we have cloned are true *TSG101* gene homologs.

Analysis of TSG101 Mutations in Human Breast Cancers. Extensive studies have shown deletion or loss of 25 heterozygosity of markers at or near the 11p15 band in a variety of human malignancies, primarily breast cancers, but also Wilms' tumor, and ovarian and testicular malignancies, suggesting that this region contains one or more tumor suppressor genes. Moreover, a region mapping 30 between 11p15.4 and 11pcen was deleted in approximately 30% of 171 sporadic breast tumors analyzed. The notion that chromosome 11 contains a tumor suppressor gene specifically implicated in the pathogenesis of human breast cancer is supported by evidence that introducing a normal chromosome 35 11 or segments of this chromosome into breast cancer cells

- 56 -

reverses their metastatic potential, as well as other properties associated with oncogenesis. The finding that homozygous inactivation of *tsg101* converts mouse fibroblasts into metastasizing cancer cells suggests that 5 this gene functions as a suppressor of malignant cell growth. To investigate the role for *TSG101* in human breast cancer, cDNA isolated from ten breast cancer cell lines and fourteen primary human breast tumors was examined specifically for mutations in *TSG101*, comparing these cDNAs 10 with cDNA obtained from three normal fibroblast cell lines, eleven tumor cell lines derived from other types of cancers, and matched normal breast tissue from the individuals that were the source of the primary breast cancers.

15 RT-PCR using primers that bracketed or were within the *TSG101* protein-coding region showed deletions in *TSG101* transcripts in five of the ten breast cancer cell lines examined. These were localized by PCR amplification using sets of primers for different segments of the gene, and the 20 fragments generated by these primers were cloned in the pCNTR plasmid vector for sequencing. The sequence of cloned *TSG101* cDNA from the independently isolated and maintained cell lines revealed a 85 bp deletion predicted to remove 28 aa (codons 5-32) and to generate, after codon 25 32, a frameshift predicted to lead to premature termination of the *TSG101* protein 10 codons later. Sequence analysis of RT-PCR fragments amplified from cell lines L7 and L8 showed deletions that remove most or all of the coiled coil domain. While a deletion of ~250 bp was observed in RT-PCR 30 cDNA generated from a fifth cell line (MDA-MB-436), using primers that amplify a 490 bp *TSG101* region located 5' to the coiled coil domain, the highly heterogeneous nature of this cell line complicated further analysis of transcripts. Three human fibroblast lines, three melanoma cell lines,

- 57 -

two Wilms' tumor cell lines, two neuroblastoma cell lines, two sarcoma cell lines, a bladder cancer cell line, and one tumor cell line from a Beckwith-Wiedemann syndrome patient showed only normal length *TSG101* transcripts by RT-PCR analysis. Normal length *TSG101* transcripts were also present in cultures of breast cancer cell lines that produced truncated *TSG101* transcripts, and variability in the ratio of normal-length and truncated *TSG101* transcripts in different experiments suggested heterogeneity in these tumor cell lines.

RT-PCR generated 1389 bp fragments amplified from each of the four breast cancer cell lines that contained specifically mapped *TSG101* deletions were cloned, and several of the clones generated from each cell were analyzed by sequencing. Comparison of the sequences from these cloned cDNAs with the sequence of *TSG101* cDNA from two normal human lines and from the melanoma cell lines revealed a C to T transition in codon 107 of transcript from L8, indicating the existence of two separate *TSG101* alleles in this breast cancer cell line.

The sequences obtained were compared with the sequences of RT-PCR products from transcripts of normal human fibroblasts (cell lines 0 and 1) and human melanoma lines (cell line 2 and 3). A point mutation in *TSG101* was identified in breast cancer cell line 8. This C to T transition results in change codon 107 from Trp to Arg. No point mutations in *TSG101* were found in an initial analysis of other tumor cell lines or in the *TSG101* sequence of melanoma cells or normal fibroblasts.

To search for mutation(s) in other *TSG101* alleles within the cell lines containing deletions in one allele of *TSG101*, the cloned 1389 bp full length RT-PCR fragments from the four breast cancers carrying *TSG101* deletions (cell lines 4, 6, 7, and 8) were sequenced. PCR amplification of the corresponding regions of *TSG101*

- 58 -

genomic DNA using primers derived from intron and exon sequences showed deletions of the expected size in the genomic DNA of all of the four breast cancer cell lines (L4, L6, L7 and L8) whose cDNA had been characterized by 5 sequencing: a 322 bp genomic PCR fragment was amplified from cell line L8, a 192 bp fragment from cell line L7 and a genomic fragment containing a deletion of about 85 bp was amplified from cell lines L4 and L6. Southern blotting of the amplified genomic fragments using a TSG101-specific DNA 10 probe confirmed that the cDNA deletions characterized had resulted from the deletion of TSG101 genomic DNA. No deletions were observed in TSG101 genomic DNA amplified from normal fibroblast cell controls.

Primary breast cancer cells produce deleted TSG101 15 transcripts. The above experiments indicate that TSG101 is mutated in a significant fraction of the breast cancer cell lines that were studies. The association of TSG101 mutations with human breast cancer implied by these cell line data was further demonstrated by analysis of fourteen 20 primary breast tumors, along with matched normal breast tissue obtained from the individuals that were the source of these tumors: six of fourteen primary breast cancers (P1, P2, P3, P4, P5 and P6) produced TSG101 transcripts containing a deletion, whereas corresponding deletions were 25 not observed in transcripts from matched normal breast tissue. Two of the five primary breast cancers that contained deleted TSG101 transcripts produced two different truncated transcript species, implying the existence of mutations in two TSG101 alleles in these cancer cells. 30 Non-deleted TSG101 transcripts were also detected in varying amounts in primary cancer specimens that produced truncated TSG101 transcripts; while some of these transcripts may result from the concurrent presence of non-tumor cells in the tissue specimens analyzed, as has been 35 observed previously, other non-deleted transcripts may be

- 59 -

derived from tumor cells and contain point mutations, as was found for breast cancer cell line L8.

Sequence analysis of truncated cDNA fragments amplified from primary breast cancers identified the TSG101 sequences that were deleted in these primary cancers. Three primary tumors, (P1, P3 and P6) had deletions in the same cDNA region. Transcripts of one primary tumor deleted in TSG101 (P4) also contained a 63 bp insertion at the deletion junction. All of the deletions that were sequenced affected the TSG101 protein-coding sequence and removed all or part of the stathmin-interacting cc2 domain.

The extraordinary conservation observed between the mouse and human TSG101 proteins is consistent with its important biological role. Both the coiled-coil and proline-rich domains are nearly identical, and the potential phosphorylation and N-glycosylation sites are completely conserved between the human and mouse protein. Chromosomal mapping of TSG101 to human chromosome 11 and mouse chromosome 7, which share conserved syntenic regions, demonstrate that the human gene and mouse genes are homologs.

Both the mouse and human TSG101 proteins contain a coiled coil domain nearly identical to one previously shown to interact with stathmin, a phosphoprotein proposed to function in the coordination and relay of diverse signals regulating cell proliferation and differentiation. The presence of multiple DNA-binding domains in the TSG101 protein and a proline-rich domain near the leucine zipper DNA binding motif of this protein indicates that the TSG101 gene product is a transcription factor, and therefore a downstream effector of stathmin action.

Deletions in TSG101 were detected in 40% of the metastatic breast cancer cell lines analysed and in the same percentage of primary tumors. A point mutation was also identified in a non-deleted TSG101 allele in one of

- 60 -

the primary tumors (MDA-MB-468; L8) and deletions in two different *TSG101* transcript species were demonstrated for two primary breast cancers by gel analysis and sequencing. The data indicate that this may represent sporadic-type 5 breast carcinomas.

All of the deletions, both in breast cancer cell lines and in primary breast tumors affected the stathmin-interacting coiled-coil domain of the *TSG101* protein. The deletions in cancer cell lines MDA-MB-435 (L7) and MDA-MB-10 468 (L8) included part or all of the coiled coil domain, whereas the deletions observed near the N-terminal end of the protein in cell lines MDA-MB-231 (L4) and MDA-MB-453 (L6) were predicted to generate a frame shift from the point of deletion and terminate the protein by a stop codon 15 10 amino acids later - prior to the coiled-coil domain. All of the mutations resulted in total or partial deletion of the cc? -coding sequence.

Interestingly, a mutation of the *p53* gene was also reported in MDA-MB-468 (L8), which contains mutations in 20 two *TSG101* alleles, suggesting the possibility that *p53* and *TSG101* have additive roles in the step-wise progression of breast-cancer to the fully malignant state.

It is noteworthy that the breast cancer cell lines having a DNA deletion that contains the *TSG101* gene have 25 also been shown to have high metastatic potential in nude mice. Line MDA-MB-435 (L7) was reported previously to contain a single copy of 11 p in many cells of the population, and FISH analysis has indicated that the majority of MDA-MB-435 cells contain only a single copy of 30 *TSG101*. This cell line was found to have a high metastatic potential and is negative for both estrogen and progesterone receptors. Introduction of a copy of normal chromosome 11 significantly suppressed this metastatic potential. These observations are consistent with the 35 finding that LOH at 11p15 in primary human breast tumors is

- 61 -

associated with poor survival after metastasis and the suggestion that LOH at 11p15 is involved in late stage tumor progression.

The TSG101 gene and the protein it encodes are 5 useful for not only the diagnosis of human breast cancer and other human cancers as well, but also for gaining an increased understanding of mechanisms of tumorigenesis.

Experimental Procedures

cDNA and Genomic DNA Cloning. The two UDG-primers 10 derived from ESTs H53754 and Z30135 were [SEQ ID NO:5] Pa-UDG (5'AGGUCAUGAUUGUGGUUUUGGAGAUG3') and [SEQ ID NO:6] Pd-UDG (5'CAUCUCCAAAUACCACAAUCAUGACCU 3'). Two UDG-primers derived from the lgt10 cloning site are [SEQ ID NO:7] Pb-UDG (5'CAUCAUCAUCAUGAGGTGGCTTATGAGTATTCTTCCAG3') and [SEQ ID 15 NO:8] Pc-UDG (5'CUACUACUACUACACCTTTGAGCAAGTTCAGCCTGGTT3'). 5' (Pc-UDG and Pd-UDG) and 3' (Pa-UDG and Pb-UDG) segments were amplified by PCR as following condition: 100 μ l final volume of 20 mM Tris-HCl pH 8.55, 3.3 mM MgCl₂, 16 mM (NH₄)₂SO₄, 150 μ g/ml BSA, 300 μ M each dNTP, 1 μ l human 20 placenta lgt10 cDNA library (titer 10⁶/ μ l, ATCC), 0.2 μ l of KlentagLA (Barnes (1994) P.N.A.S. 91:2216-2220), in a Perkin-Elmer Cetus thermal cycler for 40 cycles of: 95°C for 45 s (for denaturation), annealing and extending at 72°C for 1 min. The PCR products were visualized in 25 ethidium bromide-stained low melting agarose gels, purified and cloned into pAMP1 cloning vector (Life Technologies, Inc.). Multiple clones were isolated and both strands of the cDNA inserts were sequenced using Sequenase 2.0 (USB).

The PCR product made using primers, [SEQ ID NO:9] 5' 30 CTGATACCAGCTGGAGGTTGAGCTTTC3' (forward primer) and [SEQ ID NO:10] 5'ATTTAGCAGTCCAACATTCAAGCACAAA3' (reverse primer) were used to screen a PAC library containing human genomic DNA insert (Genome Systems, Inc.), yielding two overlapping clones, PAC1 and PAC2, each containing inserts about 150 kb

- 62 -

long. The presence of *TSG101*-specific sequences within these inserts was confirmed by Southern blotting, using a 5' fragment of human *TSG101* cDNA as probe.

Cell Lines and Cell Culture. Human breast cancer cell lines (MDA-MB-231, MDA-MB-436, MDAMB-435, MDA-MB-468, MDA-MB-157, MDA-MB-175VII, MDA-MB361, BT-483, and MCF-7), Wilms tumor cell lines (G401 and SK-NEP-1), and primary cultures of human normal fibroblast (CCD-19LuA and MRC-9) were obtained from American Type Culture Collection. Two melanoma cell lines (A375P and A375SM) were provided by I. J. Fidler. Two pliemonic sarcoma cell lines (FB 309 and FB 310), a normal fibroblast line (FB316), a bladder carcinoma (FB241), two neuroblastoma cell lines (FB616 and FB617) and a Beckman-Wiedemann syndrome tumor cell line (FB583) were obtained from the collection of Ute Francke. All cell lines were cultured in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum, 100 U/ml penicillin, and 100 µg/ml streptomycin, except for breast cancer BT-483 cells, which were cultured in RPMI-1640 medium with 20% fetal bovine serum and two Wilms tumor cell lines (G401 and SK-NEP-1), which were cultured in McCoy's 5a medium with 10% fetal bovine serum. Cell lines from the Francke collection were cultured in Mimal Essential Medium a supplemented with 10% fetal bovine serum.

Fourteen primary breast tumors and matched normal breast tissues from the same patients were obtained from Cooperative Human Tissue Network and from Biochain Institute, Inc. (San Leandro, CA).

Northern Blot Analysis. A Northern blot filter of multiple normal tissue mRNA was purchased (Clontech). Radioactively-labeled single anti-sense strand DNA probe generated from full length human *TSG101* cDNA by 40 cycles of primer extension, using [³²P]dCTP, was hybridized to the filter using standard methods. The same blot was stripped

- 63 -

and hybridized with a human β -actin probe synthesized by random priming as an internal loading control.

Somatic cell hybrids, PCR amplifications, and SSCA. The human *TSG101* gene was localized to a human chromosome 5 using a panel of 18 human X Chinese hamster hybrid cell lines derived from several independent fusion experiments (summarized in Francke et al. (1986) *Cold Spring Harb. Symp. Quant. Biol.* 2:855-866). The mouse *tsg101* gene was mapped by analyzing a mapping panel of 20 mouse X Chinese 10 hamster and two mouse X rat somatic cell hybrid lines derived from four independent fusion experiments, as described previously in Li et al. (1993) *Genomics* 18:667-672. The PCR primers used to amplify human and murine *TSG101* sequences were derived from the 3' -untranslated 15 region: the human primers were those employed to clone *TSG101* genomic DNA as described above. The murine primers were: [SEQ ID NO:11] 5'GAGACCGACCTCTCCGTAAAGCATTCTT3' (forward primer) and [SEQ ID NO:12] 5'TAGCCCAGTCAGTCCCAGCACAGCACAG 3' (reverse primer). PCR 20 conditions were 95°C, 2 min; then 35 cycles of 94°C, 30 seconds; 68°C, 30 seconds; 72°C, 1 min; followed by 72°C, 7 min. To distinguish the PCR products from human and hamster sequences in some of hybrid lines, single-strand conformation analysis (SSCA) was carried out as described 25 previously in Li et al. (1995) *Cytogenet Cell Genet* 71:301-305.

Fluorescence *in situ* hybridization. The chromosomal localization of the human *TSG101* gene was independently determined by fluorescence *in situ* hybridization (FISH). 30 Two genomic PAC1 and PAC2 clones carrying ~150 kb inserts, each containing overlapping human *TSG101* sequences, were labeled with biotin-11-dUTP by nick-translation using commercial reagents (Boehringer Mannheim). Each labeled probe was hybridized at a concentration of 200 ng/50 μ l per 35 slide to pre-treated and denatured metaphase chromosomes

- 64 -

from human lymphocytes. Hybridization, signal detection and amplification, as well as microscopy analysis and digital imaging were performed as previously described in Li et al. (1995) *Cytogenet. Cell Genet.* **68**:185-191.

5 *Human radiation hybrid mapping panel.* The Stanford G3 radiation hybrid (RH) mapping panel was purchased from Research Genetics, Inc. and was used to further define the localization of the human *TSG101* gene on human chromosome 11. This panel consists of 83 RH clones of the whole human 10 genome with a resolution of approximately 500 kb. All 83 RH cell lines were typed for the human *TSG101* gene by using primers and PCR conditions as described above. The results were sent to Stanford Human Genome Center for analysis with a software package of two-point and multipoint maximum 15 likelihood methods, described by Boehnke et al. 1991.

RT-PCR and Sequencing of cDNAs. Total RNA was isolated using RNA Stat-60 (TEL-TEST). 10 µg of total RNA was treated with 10 units of RNase-free DNase I (Boehringer Mannheim) for 10 min, extracted with phenol-chloroform 20 twice, and precipitated with ethanol. First strand cDNAs were synthesized by SuperScript II™ RNase H- reverse transcriptase (Life Technologies) using the *TSG101*-specific primer [SEQ ID NO:13] P2 (5'ATTTAGCAGTCCAACATTCAAGCACAAA3') and the human GAPDH antisense primer [SEQ ID NO:14] 25 (5'GTCTTCTGGGTGGCAGTGATGGCAT3') as a control. 1-2 µl of each product was used for PCR amplification with primer sets indicated. Primers used were [SEQ ID NO:15] P1 (5'CGGGTGTGGAGAGCCAGCTCAAGAAA3'), [SEQ ID NO:16] P3 (5'CCTTACCCACCTGGTGGTCCATATCCTG3'), [SEQ ID NO:17] P4 (5'CCTCCAGCTGGTATCAGAGAAGTCGT3') and [SEQ ID NO:18] P5 (5'CACAGTCAGACTTGTGGGGCTTATTC3'). PCR amplifications were carried out in 50 µl final volume of 20 mM Tris-HCl pH 8.55, 3.3 mM MgCl₂, 16 mM (NH₄)₂SO₄, 150 µg/ml BSA, 300 µM each dNTP, 0.2 µl of KlentagLA (Barnes, *supra.*), in a 30 Perkin-Elmer/Cetus thermal cycler for 35 cycles of 95°C 35

- 65 -

for 45 s(for denaturation), 65°C for 30s (for annealing) and extension at 72°C for 30 s to 1 min and 30s. The PCR products were visualized in ethidium bromide-stained low melting agarose gels, gel fragments were purified (Qiagen) 5 and cloned into pCNTR cloning vector (5 Prime - 3 Prime, Inc.) Multiple clones were isolated and sequenced using Sequenase 2.0 (USB).

For primary tumors and matched normal breast tissue, P1 and P4 primers were used for first PCR amplification for 10 25 cycles. The amplified products were diluted 50-fold, 1 ml of the diluted product was used for a second round PCR amplification for 30 cycles using two nested primers, P6 (SEQ ID NO:21) AGCCAGCTCAAGAAAATGGTGTCCAAG, and P7 (SEQ ID NO:22) TCACTGAGACCGGCAGTCTTCTTGCTT. The PCR products were 15 resolved in 1.5% agarose gel stained with ethidium bromide, DNA fragments were cut from gels, and purified with QIA quick gel extract kit (Qiagen). The sequence of cDNA derived from primary breast cancers and matched normal breast tissue was determined using an Applied Biosystems 20 model 310 genetic analyzer.

PCR amplification of genomic DNA. Genomic DNA was isolated by standard methods. 50 ng of RNase A treated genomic DNA was used as a template for PCR amplifications. The primers used were:

25 Cell line 8, (SEQ ID NO:23) TTCTGAAGGTTCTGTGAGACAAATAG; and (SEQ ID NO:24) CCTCCAGCTGGTATCAGAGAAG; Cell line 7, (SEQ ID NO:25) CAGTAGGGATGGCACAAATCAGCGAGGA and (SEQ ID NO:26) GGTCAGTGCCTCTACAACCCAAGTTAA; cell lines 4 and 6, (SEQ ID NO 27) CGGGTGTGGAGAGCCAGCTCAAGAAA and (SEQ 30 ID NO:28) TTTATTTTTACAAAGGTTCTGTTCTC. PCR amplifications were carried out in 50 ml final volume of 20 mM Tris-HCl pH 8.55, 3.3 mM MgCl₂, 16 mM (NH₄)₂SO₄, 150 mg/ml BSA, 300 mM each dNTP, 0.2 ml of KlentaqLA, in a Perkin Elmer/Cetus thermal cycler for 40 cycles at 95°C for

- 66 -

45 s, 65°C for 30 s, and extension at 72°C for 30 s to 1 min.

All publications and patent applications cited in this specification are herein incorporated by reference as 5 if each individual publication or patent application were specifically and individually indicated to be incorporated by reference.

Although the foregoing invention has been described in some detail by way of illustration and example for 10 purposes of clarity of understanding, it will be readily apparent to those of ordinary skill in the art in light of the teachings of this invention that certain changes and modifications may be made thereto without departing from the spirit or scope of the appended claims.

- 67 -

SEQUENCE LISTING

(1) GENERAL INFORMATION:

5 (i) APPLICANT: THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR
UNIVERSITY

(ii) TITLE OF INVENTION: MAMMALIAN TUMOR SUSCEPTIBILITY GENES AND
THEIR USES

10 (iii) NUMBER OF SEQUENCES: 28

10 (iv) CORRESPONDENCE ADDRESS:
FISH AND RICHARDSON, P.C.
2200 SAND HILL ROAD
MENLO PARK, CA 94025

15 (v) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: Floppy disk
(B) COMPUTER: IBM PC compatible
(C) OPERATING SYSTEM: PC-DOS/MS-DOS
(D) SOFTWARE: PatentIn Release #1.0, Version #1.30

20 (vi) CURRENT APPLICATION DATA:
(A) APPLICATION NUMBER:
(B) FILING DATE:
(C) CLASSIFICATION:

25 (viii) ATTORNEY/AGENT INFORMATION:
(A) NAME: SHERWOOD, Pamela J.
(B) REGISTRATION NUMBER: 36,677

(ix) TELECOMMUNICATION INFORMATION:
(A) TELEPHONE: 415-322-5070
(B) TELEFAX: 415-854-0875

(2) INFORMATION FOR SEQ ID NO:1:

30 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 1448 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

35 (ii) MOLECULE TYPE: cDNA

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 61..1203

40 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

CCCCCTCTGCC TGTGGGGACCG GAGGAGCGCG CCATGGCTGT CCGAGAGTCA GCTGAAGAAG
60

ATG ATG TCC AAG TAC AAA TAT AGA GAT CTA ACC GTC CGT CAA ACT GTC
108

- 68 -

Met	Met	Ser	Lys	Tyr	Lys	Tyr	Arg	Asp	Leu	Thr	Val	Arg	Gln	Thr	Val	
1					5				10					15		
AAT GTC ATC GCT ATG TAC AAA GAT CTC AAA CCT GTA TTG GAT TCA TAT																
156																
5	Asn	Val	Ile	Ala	Met	Tyr	Lys	Asp	Leu	Lys	Pro	Val	Leu	Asp	Ser	Tyr
					20				25					30		
GTT TTT AAT GAT GGC AGT TCC AGG GAG CTG GTG AAC CTC ACT GGT ACA																
204																
10	Val	Phe	Asn	Asp	Gly	Ser	Ser	Arg	Glu	Leu	Val	Asn	Leu	Thr	Gly	Thr
					35				40					45		
ATC CCA GTG CGT TAT CGA GGT AAT ATA TAT AAT ATT CCA ATA TGC CTG																
252																
	Ile	Pro	Val	Arg	Tyr	Arg	Gly	Asn	Ile	Tyr	Asn	Ile	Pro	Ile	Cys	Leu
					50				55					60		
15	TGG	CTG	CTG	GAC	ACA	TAC	CCA	TAT	AAC	CCC	CCT	ATC	TGT	TTT	GTT	AAG
					300											
	Trp	Leu	Leu	Asp	Thr	Tyr	Pro	Tyr	Asn	Pro	Pro	Ile	Cys	Phe	Val	Lys
					65				70					75	80	
20	CCT	ACT	AGT	TCA	ATG	ACT	ATT	AAA	ACA	GGA	AAG	CAT	GTG	GAT	GCA	AAT
					348											
	Pro	Thr	Ser	Ser	Met	Thr	Ile	Lys	Thr	Gly	Lys	His	Val	Asp	Ala	Asn
					85				90					95		
GGG AAA ATC TAC CTA CCT TAT CTA CAT GAC TGG AAA CAT CCA CGG TCA																
396																
25	Gly	Lys	Ile	Tyr	Leu	Pro	Tyr	Leu	His	Asp	Trp	Lys	His	Pro	Arg	Ser
					100				105					110		
GAG TTG CTG GAG CTT ATT CAA ATC ATG ATT GTG ATA TTT GGA GAG GAG																
444																
30	Glu	Leu	Leu	Glu	Leu	Ile	Gln	Ile	Met	Ile	Val	Ile	Phe	Gly	Glu	Glu
					115				120					125		
CCT CCA GTG TTC TCC CGG CCT ACT GTT TCT GCA TCC TAC CCA CCA TAC																
492																
	Pro	Pro	Val	Phe	Ser	Arg	Pro	Thr	Val	Ser	Ala	Ser	Tyr	Pro	Pro	Tyr
					130				135					140		
35	ACA	GCA	ACA	GGG	CCA	CCA	AAT	ACC	TCC	TAC	ATG	CCA	GGC	ATG	CCA	AGT
					540											
	Thr	Ala	Thr	Gly	Pro	Pro	Asn	Thr	Ser	Tyr	Met	Pro	Gly	Met	Pro	Ser
					145				150				155		160	
40	GGA	ATC	TCT	GCA	TAT	CCA	TCT	GGA	TAC	CCT	CCC	AAC	CCC	AGT	GGT	TAT
					588											
	Gly	Ile	Ser	Ala	Tyr	Pro	Ser	Gly	Tyr	Pro	Pro	Asn	Pro	Ser	Gly	Tyr
					165				170					175		
CCT GGC TGT CCT TAC CCA CCT GCT GGC CCA TAC CCT GCC ACA ACA AGC																
636																
45	Pro	Gly	Cys	Pro	Tyr	Pro	Pro	Ala	Gly	Pro	Tyr	Pro	Ala	Thr	Thr	Ser
					180				185					190		
TCA CAG TAC CCT TCC CAG CCT GTG ACC ACT GTT GGT CCC AGC AGA																
684																
50	Ser	Gln	Tyr	Pro	Ser	Gln	Pro	Pro	Val	Thr	Thr	Val	Gly	Pro	Ser	Arg
					195				200					205		

- 69 -

GAT GGC ACA ATC AGT GAG GAC ACT ATC CGT GCA TCT CTC ATC TCA GCA
 732
 Asp Gly Thr Ile Ser Glu Asp Thr Ile Arg Ala Ser Leu Ile Ser Ala
 210 215 220
 5 GTC AGT GAC AAA CTG AGA TGG CGG ATG AAG GAG GAA ATG GAT GGT GCC
 780
 Val Ser Asp Lys Leu Arg Trp Arg Met Lys Glu Glu Met Asp Gly Ala
 225 230 235 240
 10 CAG GCA GAG CTT AAT GCC TTG AAA CGA ACA GAG GAA GAT CTG AAA AAA
 828
 Gln Ala Glu Leu Asn Ala Leu Lys Arg Thr Glu Glu Asp Leu Lys Lys
 245 250 255
 15 GGC CAC CAG AAA CTG GAA GAG ATG GTC ACC CGC TTA GAT CAA GAA GTA
 876
 Gly His Gln Lys Leu Glu Glu Met Val Thr Arg Leu Asp Gln Glu Val
 260 265 270
 20 GCT GAA GTT GAT AAA AAC ATA GAA CTT TTG AAA AAG AAG GAT GAA GAA
 924
 Ala Glu Val Asp Lys Asn Ile Glu Leu Leu Lys Lys Asp Glu Glu
 275 280 285
 25 CTA AGT TCT GCT CTG GAG AAA ATG GAA AAT CAA TCT GAA AAT AAT GAT
 972
 Leu Ser Ser Ala Leu Glu Lys Met Glu Asn Gln Ser Glu Asn Asn Asp
 290 295 300
 30 ATT GAT GAA GTT ATC ATT CCC ACA GCC CCA CTG TAT AAA CAG ATT CTA
 1020
 Ile Asp Glu Val Ile Ile Pro Thr Ala Pro Leu Tyr Lys Gln Ile Leu
 305 310 315 320
 35 AAT CTG TAT GCA GAG GAA AAT GCT ATT GAA GAC ACT ATC TTT TAC CTT
 1068
 Asn Leu Tyr Ala Glu Glu Asn Ala Ile Glu Asp Thr Ile Phe Tyr Leu
 325 330 335
 40 GGA GAA GCT TTG CGG CGG GGA GTC ATA GAC CTG GAT GTG TTC CTG AAA
 1116
 Gly Glu Ala Leu Arg Arg Gly Val Ile Asp Leu Asp Val Phe Leu Lys
 340 345 350
 CAC GTC CGC CTC CTG TCC CGT AAA CAG TTC CAG CTA AGG GCA CTA ATG
 1164
 His Val Arg Leu Leu Ser Arg Lys Gln Phe Gln Leu Arg Ala Leu Met
 355 360 365
 45 CAA AAG GCA AGG AAG ACT GCG GGC CTT AGT GAC CTC TAC TGACATGTGC
 1213
 Gln Lys Ala Arg Lys Thr Ala Gly Leu Ser Asp Leu Tyr
 370 375 380
 50 TGTCAGCTGG AGACCGACCT CTCCGTAAAG CATTCTTTTC TTCTTCTTT TCTCATCAGT
 1273
 AGAACCCACA ATAAGTTATT GCAGTTTATC ATTCAAGTGT TAAATATTTT GAATCAATAA
 1333
 TATATTTCT GTTCCTTG GGTAAAAACT GGCTTTATT AATGCACTTT CTACCCCTCTG
 1393

- 70 -

TAAGCGTCTG TGCTGTGCTG GGACTGACTG GGCTAAATAA AATTGTTGC ATAAA
1448

(2) INFORMATION FOR SEQ ID NO:2:

5 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 381 amino acids
 (B) TYPE: amino acid
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

10 Met Met Ser Lys Tyr Lys Tyr Arg Asp Leu Thr Val Arg Gln Thr Val
 1 5 10 15

Asn Val Ile Ala Met Tyr Lys Asp Leu Lys Pro Val Leu Asp Ser Tyr
 20 25 30

15 Val Phe Asn Asp Gly Ser Ser Arg Glu Leu Val Asn Leu Thr Gly Thr
 35 40 45

Ile Pro Val Arg Tyr Arg Gly Asn Ile Tyr Asn Ile Pro Ile Cys Leu
 50 55 60

Trp Leu Leu Asp Thr Tyr Pro Tyr Asn Pro Pro Ile Cys Phe Val Lys
 65 70 75 80

20 Pro Thr Ser Ser Met Thr Ile Lys Thr Gly Lys His Val Asp Ala Asn
 85 90 95

Gly Lys Ile Tyr Leu Pro Tyr Leu His Asp Trp Lys His Pro Arg Ser
 100 105 110

25 Glu Leu Leu Glu Leu Ile Gln Ile Met Ile Val Ile Phe Gly Glu Glu
 115 120 125

Pro Pro Val Phe Ser Arg Pro Thr Val Ser Ala Ser Tyr Pro Pro Tyr
 130 135 140

Thr Ala Thr Gly Pro Pro Asn Thr Ser Tyr Met Pro Gly Met Pro Ser
 145 150 155 160

30 Gly Ile Ser Ala Tyr Pro Ser Gly Tyr Pro Pro Asn Pro Ser Gly Tyr
 165 170 175

Pro Gly Cys Pro Tyr Pro Pro Ala Gly Pro Tyr Pro Ala Thr Thr Ser
 180 185 190

35 Ser Gln Tyr Pro Ser Gln Pro Pro Val Thr Thr Val Gly Pro Ser Arg
 195 200 205

Asp Gly Thr Ile Ser Glu Asp Thr Ile Arg Ala Ser Leu Ile Ser Ala
 210 215 220

Val Ser Asp Lys Leu Arg Trp Arg Met Lys Glu Glu Met Asp Gly Ala
 225 230 235 240

40 Gln Ala Glu Leu Asn Ala Leu Lys Arg Thr Glu Glu Asp Leu Lys Lys
 245 250 255

- 71 -

Gly His Gln Lys Leu Glu Glu Met Val Thr Arg Leu Asp Gln Glu Val
 260 265 270

Ala Glu Val Asp Lys Asn Ile Glu Leu Leu Lys Lys Asp Glu Glu
 275 280 285

5 Leu Ser Ser Ala Leu Glu Lys Met Glu Asn Gln Ser Glu Asn Asn Asp
 290 295 300

Ile Asp Glu Val Ile Ile Pro Thr Ala Pro Leu Tyr Lys Gln Ile Leu
 305 310 315 320

10 Asn Leu Tyr Ala Glu Glu Asn Ala Ile Glu Asp Thr Ile Phe Tyr Leu
 325 330 335

Gly Glu Ala Leu Arg Arg Gly Val Ile Asp Leu Asp Val Phe Leu Lys
 340 345 350

His Val Arg Leu Leu Ser Arg Lys Gln Phe Gln Leu Arg Ala Leu Met
 355 360 365

15 Gln Lys Ala Arg Lys Thr Ala Gly Leu Ser Asp Leu Tyr
 370 375 380

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:

20 (A) LENGTH: 1494 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

25 (ix) FEATURE:
 (A) NAME/KEY: CDS
 (B) LOCATION: 120..1259

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

GAAGGGGGTG TCGGATTGTG TGGGACGGTC TGGGGCAGCC ACAGCGGCTG ACCNCNTNGC
 60

30 CTGCGGGGAA GGGAGTCGCC AGGGCCCGTC ATCGGGTGTG GGAGAGCCAG CTCAAGAAAA
 120

TGGTGTCCAA GTACAAATAC AGAGACCTAA CTGTACGTGA AACTGTCAAT GTTATTACTC
 180

TATACAAAGA TCTCAAACCT GTTTGGATT CATATGTTT TAACGATGGC AGTTCCAGGG
 35 240

AACTAATGAA CCTCACTGGA ACAATCCCTG TGCCTTATAG AGGTAATACA TACAATATTC
 300

CAATATGCCT ATGGCTACTG GACACATACC CATATAATCC CCCTATCTGT TTTGTTAAGC
 360

40 CTACTAGTTC AATGACTATT AAAACAGGAA AGCATGTTGA TGCAAATGGG AAGATATATC
 420

- 72 -

TTCCCTTATCT ACATGAATGG AAACACCCAC AGTCAGACTT GTTGGGGCTT ATTCAAGGTCA
 480
 TGATTGTGGT ATTTGGAGAT GAACCTCCAG TCTTCTCTCG TCCTATTCG GCATCCTATC
 540
 5 CGCCATACCA GGCAACGGGG CCACCAAATA CTTCCTACAT GCCAGGCATG CCAGGTGGAA
 600
 TCTCTCCATA CCCATCCGGA TACCCCTCCA ATCCCAGTGG TTACCCAGGC TGTCCTTACC
 660
 10 CACCTGGTGG TCCATATCCT GCCACAAACAA GTTCTCAGTA CCCTTCTCAG CCTCCTGTGA
 720
 CCACTGTTGG TCCCAGTAGG GATGGCACAA TCAGCGAGGA CACCATCCGA GCCTCTCTCA
 780
 TCTCTGCGGT CAGTGACAAA CTGAGATGGC GGATGAAGGA GGAAATGGAT CGTCCCCAGG
 840
 15 CAGAGCTCAA TGCCTTGAAA CGAACAGAAG AAGACCTGAA AAAGGGTCAC CAGAAACTGG
 900
 AAGAGATGGT TACCCGTTA GATCAAGAAG TAGCCGAGGT TGATAAAAAC ATAGAACTTT
 960
 20 TGAAAAAGAA GGATGAAGAA CTCAGTTCTG CTCTGGAAAA AATGGAAAAT CAGTCTGAAA
 1020
 ACAATGATAT CGATGAAGTT ATCATTCCA CAGCTCCCTT ATACAAACAG ATCCTGAATC
 1080
 TGTATGCAGA AGAAAACGCT ATTGAAGACA CTATCTTTA CTTGGGAGAA GCCTTGAGAA
 1140
 25 GGGGCGTGAT AGACCTGGAT GTCTTCCTGA AGCATGTACG TCTTCTGTCC CGTAAACAGT
 1200
 TCCAGCTGAG GGCACTAATG CAAAAACCAA GAAAGACTGC CGGTCTCAGT GACCTCTACT
 1260
 30 GACTTCTCTG ATACCAGCTG GAGGTTGAGC TCTTCTTAAA GTATTCTTCT CTTCTTTA
 1320
 TCAGTAGGTG CCCAGAATAA GTTATTGCAG TTTATCATTC AAGTGTAAAA TATTTGAAT
 1380
 CAATAATATA TTTTCTGTTT TCTTTGGTA AAGACTGGCT TTTATTAATG CACTTTCTAT
 1440
 35 CCTCTGTAAA CTTTTGTGC TGAATGTTGG GACTGCTAAA TAAAATTTGT TTTT
 1494

(2) INFORMATION FOR SEQ ID NO:4:

40 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 380 amino acids
 (B) TYPE: amino acid
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

- 73 -

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

Met Val Ser Lys Tyr Lys Tyr Arg Asp Leu Thr Val Arg Glu Thr Val
 1 5 10 15

Asn Val Ile Thr Leu Tyr Lys Asp Leu Lys Pro Val Leu Asp Ser Tyr
 5 20 25 30

Val Phe Asn Asp Gly Ser Ser Arg Glu Leu Met Asn Leu Thr Gly Thr
 35 40 45

Ile Pro Val Pro Tyr Arg Gly Asn Thr Tyr Asn Ile Pro Ile Cys Leu
 50 55 60

10 Trp Leu Leu Asp Thr Tyr Pro Tyr Asn Pro Pro Ile Cys Phe Val Lys
 65 70 75 80

Pro Thr Ser Ser Met Thr Ile Lys Thr Gly Lys His Val Asp Ala Asn
 85 90 95

Gly Lys Ile Tyr Leu Pro Tyr Leu His Glu Trp Lys His Pro Gln Ser
 15 100 105 110

Asp Leu Leu Gly Leu Ile Gln Val Met Ile Val Val Phe Gly Asp Glu
 115 120 125

20 Pro Pro Val Phe Ser Arg Pro Ile Ser Ala Ser Tyr Pro Pro Tyr Gln
 130 135 140

Ala Thr Gly Pro Pro Asn Thr Ser Tyr Met Pro Gly Met Pro Gly Gly
 145 150 155 160

Ile Ser Pro Tyr Pro Ser Gly Tyr Pro Pro Asn Pro Ser Gly Tyr Pro
 165 170 175

25 Gly Cys Pro Tyr Pro Pro Gly Gly Pro Tyr Pro Ala Thr Thr Ser Ser
 180 185 190

Gln Tyr Pro Ser Gln Pro Pro Val Thr Thr Val Gly Pro Ser Arg Asp
 195 200 205

30 Gly Thr Ile Ser Glu Asp Thr Ile Arg Ala Ser Leu Ile Ser Ala Val
 210 215 220

Ser Asp Lys Leu Arg Trp Arg Met Lys Glu Glu Met Asp Arg Ala Gln
 225 230 235 240

Ala Glu Leu Asn Ala Leu Lys Arg Thr Glu Glu Asp Leu Lys Lys Gly
 245 250 255

35 His Gln Lys Leu Glu Glu Met Val Thr Arg Leu Asp Gln Glu Val Ala
 260 265 270

Glu Val Asp Lys Asn Ile Glu Leu Leu Lys Lys Asp Glu Glu Leu
 275 280 285

40 Ser Ser Ala Leu Glu Lys Met Glu Asn Gln Ser Glu Asn Asn Asp Ile
 290 295 300

Asp Glu Val Ile Ile Pro Thr Ala Pro Leu Tyr Lys Gln Ile Leu Asn
 305 310 315 320

Leu Tyr Ala Glu Glu Asn Ala Ile Glu Asp Thr Ile Phe Tyr Leu Gly

- 74 -

325

330

335

10 Glu Ala Leu Arg Arg Gly Val Ile Asp Leu Asp Val Phe Leu Lys His
340 345 350

5 Val Arg Leu Leu Ser Arg Lys Gln Phe Gln Leu Arg Ala Leu Met Gln
355 360 365

Lys Ala Arg Lys Thr Ala Gly Leu Ser Asp Leu Tyr
370 375 380

(2) INFORMATION FOR SEQ ID NO:5:

10 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 9 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

15 (ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

His Thr His Leu Ala Met Asx Asp Ala
1 5

(2) INFORMATION FOR SEQ ID NO:6:

20 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 10 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

25 (ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

Phe Xaa Asn Gly Ala Leu Glx Cys Tyr Ser
1 5 10

(2) INFORMATION FOR SEQ ID NO:7:

30 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 27 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

35 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

AGGUCAUGAU UGUGGUAUUU GGAGAUG
27

- 75 -

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 27 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

5 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

10 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

CAUCUCCAAA UACCACAAUC AUGACCU
27

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 39 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

15 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

20 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

CAUCAUCAUC AUGAGGTGGC TTATGAGTAT TTCTTCCAG
39

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 39 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

25 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

CUACUACUAC UACACCTTT GAGCAAGTTC AGCCTGGTT
39

35 (2) INFORMATION FOR SEQ ID NO:11:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 28 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single

- 76 -

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

5 CTGATACCAAG CTGGAGGTTG AGCTCTTC
28

(2) INFORMATION FOR SEQ ID NO:12:

10 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 28 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear15 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

ATTTAGCAGT CCCAACATTC AGCACAAA
28

(2) INFORMATION FOR SEQ ID NO:13:

20 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 28 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear25 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

GAGACCGACC TCTCCGTAAA GCATTCTT
28

30 (2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 28 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
35 (D) TOPOLOGY: linear(ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

40 TAGCCCAGTC AGTCCCAGCA CAGCACAG
28

- 77 -

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 28 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

5 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:

10 ATTTAGCAGT CCCAACATTC AGCACAAA
28

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 25 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

15 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

20 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

GTCTTCTGGG TGGCAGTGAT GGCAT
25

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 27 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

25 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:

CGGGTGTCCG AGAGCCAGCT CAAGAAA
27

(2) INFORMATION FOR SEQ ID NO:18:

35 (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 28 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single

- 78 -

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

5 CCTTACCCAC CTGGTGGTCC ATATCCTG
28

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 26 base pairs
10 (B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

15 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

CCTCCAGCTG GTATCAGAGA AGTCGT
26

(2) INFORMATION FOR SEQ ID NO:20:

20 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 27 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

25 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

CACAGTCAGA CTTGTTGGGG CTTATTC
27

(2) INFORMATION FOR SEQ ID NO:21:

30 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 27 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

35 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:21:

- 79 -

AGCCAGCTCA AGAAAATGGT GTCCAAG
27

(2) INFORMATION FOR SEQ ID NO:22:

5 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 28 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

10 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:22:

TCACTGAGAC CGGCAGTCTT TCTTGCTT
28

(2) INFORMATION FOR SEQ ID NO:23:

15 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 27 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

20 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

TTCTGAAGGT TCCTGTGAGA CAAATAG

27

(2) INFORMATION FOR SEQ ID NO:24:

25 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 22 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

30 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:

CCTCCAGCTG GTATCAGAGA AG

22

(2) INFORMATION FOR SEQ ID NO:25:

35 (i) SEQUENCE CHARACTERISTICS:

- 80 -

- (A) LENGTH: 27 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

5 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:

CAGTAGGGAT GGCACAATCA GCGAGGA

27

(2) INFORMATION FOR SEQ ID NO:26:

10 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 27 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

15 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

GGTCAGTGCC TCTACAACCC AAGTTAA

27

(2) INFORMATION FOR SEQ ID NO:27:

20 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 27 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

25 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:27:

CGGGTGTGCGG AGAGCCAGCT CAAGAAA

27

(2) INFORMATION FOR SEQ ID NO:28:

30 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 29 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

35 (ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "Primer"

- 81 -

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:28:

TTTATTTTTT TACAAAGGTT TCTGTTCTC

29

- 82 -

WHAT IS CLAIMED IS:

1. An isolated nucleic acid encoding a TSG101 protein, or fragment of at least about 100 nt in length thereof, as other than an intact chromosome.
- 5 2. An isolated nucleic acid according to Claim 1, wherein said TSG101 protein is a mammalian protein.
3. An isolated nucleic acid according to Claim 2, wherein said nucleic acid comprises an oncogenic mutation.
- 10 4. An isolated nucleic acid according to Claim 3, wherein said oncogenic mutation disrupts the coiled coil domain.
- 15 5. An expression cassette comprising a transcriptional initiation region functional in an expression host, a nucleic acid having a sequence of the isolated nucleic acid according to Claim 1 under the transcriptional regulation of said transcriptional initiation region, and a transcriptional termination region functional in said expression host.
- 20 6. A cell comprising an expression cassette according to Claim 5 as part of an extrachromosomal element or integrated into the genome of a host cell as a result of introduction of said expression cassette into said host cell and the cellular progeny of said host cell.
- 25 7. A method for producing TSG101 protein, said method comprising:
 - growing a cell according to Claim 6, whereby said TSG101 protein is expressed; and
 - isolating said TSG101 protein free of other proteins.

- 83 -

8. A purified polypeptide composition comprising at least 50 weight % of the protein present as a *TSG101* protein or a fragment thereof.

9. A purified polypeptide composition according to 5 Claim 8, wherein said *TSG101* protein is a mammalian protein.

10. A monoclonal antibody binding specifically to a *TSG101* protein.

11. A method for characterizing the phenotype of a 10 tumor, the method comprising:

detecting the presence of an oncogenic mutation in *TSG101* in said tumor,

wherein the presence of said oncogenic mutation indicates that said tumor has a *TSG101*-associated 15 phenotype.

12. A method according to Claim 19, wherein said carcinoma is a breast carcinoma.

13. A method for inactivating multiple copies of a gene at an expressed unselected chromosomal locus of eukaryotic 20 cells, comprising;

introducing into said eukaryotic cells a knockout DNA construct to produce a genetically modified cell mixture, said knockout DNA construct comprising at least (i) an agent regulated promoter ("TF promoter") oriented for RNA 25 transcription in the opposite direction to (ii) a first positive selection marker coding sequence located 5' of said TF promoter, wherein a transactivation factor is provided extrinsic to said eukaryotic cells or intrinsic to said cells by introducing a transactivation DNA construct, 30 said transactivation DNA construct comprising at least (i)

- 84 -

a gene sequence for a second positive selection marker, and
(ii) a gene sequence for said transactivation factor which
binds to said transcription initiation region of said
knockout construct to initiate RNA transcription, whereby
5 an antisense RNA is produced of the sequence of integration
of knockout construct locus; and

growing said genetically modified cell mixture in a
selective medium to obtain selected genetically modified
cells, said genetically modified selected cell being
10 characterized by (i) expression of said first positive
selection marker coding sequence resulting from said
knockout DNA construct being integrated downstream of a
promoter for said gene at said random chromosomal locus and
under its transcriptional regulatory control, and, in the
15 presence of said agent or when present, (ii) expression of
said second positive selection marker gene sequence
resulting in production of transactivator factor, wherein
the first copy of said gene at said chromosomal locus is
inactivated by integration of said knockout construct
20 downstream of said promoter and any other similar genes are
inactivated by said antisense RNA.

14. The method of claim 13, further including the steps
of assaying said genetically modified for a change in cell
phenotype associated with inactivating multiple copies of
25 said gene, and determining the gene at said locus.

15. The method of claim 13, wherein said introducing
step includes introducing said knockout and transactivation
DNA constructs successively so that said knockout construct
is introduced first to produce first genetically modified
30 cells and said transactivation DNA construct is introduced
later to produce second genetically modified cells.

- 85 -

16. The method of claim 15, wherein said knockout construct and said transactivation DNA construct comprise first and second positive selection markers, respectively, and said growing step includes growing said genetically modified cells which comprise said knockout construct in a first selective medium to obtain first selected cells expressing said first selection marker and growing said second genetically modified cells in a second selective medium to obtain second selected cells expressing both positive selection marker sequences.

17. The method of claim 13, wherein said knockout DNA construct further includes a splice acceptor sequence which is 5' in relationship to said positive selection marker coding region sequence.

15 18. The method of claim 17, wherein said splice acceptor sequence is 3' in relationship to said TF promoter.

19. The method of claim 13, wherein said transactivation factor includes a transcription activation domain and a DNA-binding domain, and said TF promoter includes a 20 promoter sequence linked to multiple copies of a sequence which binds said DNA-binding domain, said DNA binding domain being exogenous to said eukaryotic cells.

20. The method of claim 19, wherein said promoter sequence comprises a domain derived from 25 a viral transcription regulatory protein gene and said DNA-binding domain is derived from the lac repressor protein, and said sequence which binds said DNA-binding domain includes multiple copies of the lac operator sequence.

21. A knockout DNA construct sequence comprising a 30 promoterless positive selection marker coding sequence and

- 86 -

a TF promoter responsive to a transactivation factor located in the direction of transcription of said coding sequence 5' of said coding sequence and oriented for transcription in the direction opposite said coding sequence.

22. A transactivation DNA construct sequence comprising a gene sequence for a transactivation factor, a gene sequence for a positive selection marker, and a gene sequence for a negative selection marker, said gene sequence for a transactivation factor and said gene sequence for said negative selection marker being delimited by two site-specific recombination sites.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US96/18828

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : C12Q 1/68; C12N 15/00; C07H 21/04
US CL : 435/6, 172.3, 320.1; 536/24.5

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6, 172.3, 320.1; 536/24.5

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, Medline, Biosis
search terms: antisense, promoter, upstream, readthrough, flanking, cre, lox, site specific recombination,

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Maucuer, et al. Stathmin interaction with a putative kinase and coiled-coil-forming protein domains. Proc. Natl. Acad. Sci. USA April 1995, Vol. 92, pages 3100-3104, especially page 3103.	1, 2 -----
Y	Li et al. tsg 101: A novel tumor susceptibility gene isolated by controlled homozygous functional knockout of allelic loci in mammalian cells. Cell. 03 May 1996. Vol. 85. pages 319-329, especially page 320.	5-10
X, P		1-22
Y	F. Ausubel et al., 'Current Protocols in Molecular Biology', published 1994 by John Wiley and Sons, Inc. see pages 11.4.1-11.12.1, 16.0.5-16.1.3, and 16.11.19-16.12.6.	5-10

 Further documents are listed in the continuation of Box C. See patent family annex.

- * Special categories of cited documents:
- *A* document defining the general state of the art which is not considered to be of particular relevance
- *E* earlier document published on or after the international filing date
- *L* document which may throw doubt on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- *O* document referring to an oral disclosure, use, exhibition or other means
- *P* document published prior to the international filing date but later than the priority date claimed
- *T* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- *X* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- *Y* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
- *Z* document member of the same patent family

Date of the actual completion of the international search

11 FEBRUARY 1997

Date of mailing of the international search report

19 MAR 1997

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized Officer

JOHN S. BRUSCA

Telephone No. 308-0196

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US96/18828

Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)

This international report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.
3. As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

The additional search fees were accompanied by the applicant's protest.

No protest accompanied the payment of additional search fees.